Requested Patent: JP10074250A

Title:

METHOD AND APPARATUS FOR IMAGE BASED DOCUMENT PROCESSING ;

Abstracted Patent: US5943443 ;

Publication Date: 1999-08-24 ;

Inventor(s): ITONORI KATSUHIKO (JP); OZAKI MASAHARU (JP) ;

Applicant(s): FUJI XEROX CO LTD (JP) ;

Application Number: US19970880399 19970623 ;

Priority Number(s): JP19960166147 19960626; JP19960274732 19961017 ;

IPC Classification: G06K9/62; G06K9/72; G06K9/54; G06K9/60 ;

Equivalents: JP2973944B2 ;

ABSTRACT:

The present invention provides a document processing apparatus, document processing method and a storage medium for storing thereof on purpose to offer document filing in which document can be registered with a little computation cost and with high speed, and retrieval can be performed with little oversight. In the document processing apparatus, a similar character classifying element classifies characters in a document image into similar character categories in advance and stores the classified categories together with their representative image features. When the document image is registered, a pseudo character recognizing element executes, without identifying each character in the text region, classification into character categories based on the image features less than those used in the ordinary character recognition and stores the category strings generated by identifying each character with the inputted image. In retrieval, a retrieval executing element converts each character in the retrieval keyword into nearest category, and retrieves a document including the converted category string as a part as a result of retrieval.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平10－74250

(43)公開日　平成10年(1998)３月17日

(51)Int.Cl.⁶		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｔ	1/00			Ｇ０６Ｆ	15/62	３３０Ａ
Ｇ０６Ｆ	17/21				15/20	５７０Ｎ
	17/30				15/403	３１０Ｃ
						３３０Ｂ

審査請求　有　請求項の数15　ＯＬ　（全 34 頁）

(21)出願番号　特願平8－274732

(22)出願日　平成８年(1996)10月17日

(31)優先権主張番号　特願平8－166147

(32)優先日　平８(1996)６月26日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　000005496
富士ゼロックス株式会社
東京都港区赤坂二丁目17番22号

(72)発明者　糸乗　勝彦
神奈川県足柄上郡中井町境430　グリーンテクなかい富士ゼロックス株式会社内

(72)発明者　尾崎　正治
神奈川県足柄上郡中井町境430　グリーンテクなかい富士ゼロックス株式会社内

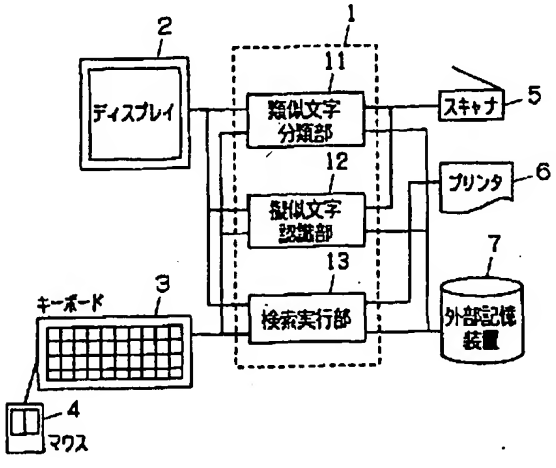
(74)代理人　弁理士　石井　康夫　（外１名）

(54)【発明の名称】　文書処理装置、文書処理方法、および記憶媒体

(57)【要約】

【課題】　文書登録時に少ない計算機パワーでしかも高速に登録処理が行なえるとともに、検索時には漏れの少ない検索を実現することのできる文書ファイリングを提供する。

【解決手段】　類似文字分類部１１において、文字画像をその画像特徴をもとに類似した文字ごとに類似文字カテゴリにあらかじめ分類し、分類されたカテゴリをその代表となる画像特徴とともに記憶しておく。文書画像登録時には、擬似文字認識部１２において、そのテキスト領域の各文字を認識せずに、文字認識を行なうよりも少ない画像特徴をもとに文字カテゴリに分類し、各文字ごとに識別されたカテゴリ列を入力画像とともに記憶しておく。検索時には検索実行部１３において、検索キーワードの各文字を対応するカテゴリに変換し、変換されたカテゴリ列を一部に含む文書を検索結果として取り出す。



【特許請求の範囲】

【請求項1】　文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリを前記画像特徴と対応づけて記憶しておく文字カテゴリ記憶手段と、入力された文書画像中の文字ごとに画像を切り出すテキスト領域抽出手段と、該テキスト領域抽出手段によって切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ記憶手段に記憶されているカテゴリに分類する擬似文字認識手段と、該擬似文字認識手段によって分類された前記各文字画像のカテゴリを前記入力された文書画像と対応づけて記憶しておく擬似文字認識結果記憶手段と、検索時に入力された検索式中のキーワードの各文字を前記文字カテゴリ記憶手段に記憶されている対応するカテゴリに変換するキーワード変換手段と、該キーワード変換手段によってカテゴリに変換された検索式を満たすカテゴリを有する文書画像を前記擬似文字認識結果記憶手段から取り出す文書検索手段を具備することを特徴とする文書処理装置。

【請求項2】　前記文字カテゴリ記憶手段に記憶されているカテゴリは、文字画像の特徴ベクトルによるクラスタリングによって分類を行なったものであることを特徴とする請求項1に記載の文書処理装置。

【請求項3】　前記擬似文字認識結果記憶手段内の文書画像に対応づけて記憶しているカテゴリは、文書画像内において隣り合う2つの文字画像のカテゴリをキーとして該キーが出現する文書の識別子を記憶するバイグラムテーブルとして記憶されており、前記文書検索手段は、前記キーワード変換手段によって変換されたカテゴリを前記バイグラムテーブルから検索することを特徴とする請求項1に記載の文書処理装置。

【請求項4】　前記文字カテゴリ記憶手段は、1つの文字を複数のカテゴリに記憶している場合があり、前記キーワード変換手段は、1つの検索キーワードに対して前記文字カテゴリ記憶手段内に記憶されているすべてのカテゴリに変換することを特徴とする請求項1に記載の文書処理装置。

【請求項5】　前記文字カテゴリ記憶手段は、1つの文字を複数のカテゴリに記憶している場合があるとともにそれぞれのカテゴリに分類される確率を記憶し、前記文書検索手段は、前記文字カテゴリ記憶手段内の確率に応じて文書画像を前記擬似文字認識結果記憶手段から取り出すことを特徴とする請求項1に記載の文書処理装置。

【請求項6】　前記テキスト領域抽出手段は、複数の文字切り出し解釈が存在する場合は該解釈すべてについて切り出しを行ない、前記擬似文字認識手段は、前記テキスト領域抽出手段により切り出されたすべての切り出し結果に対してカテゴリに分類し、擬似文字認識結果記憶手段は、前記擬似文字認識手段により分類されたすべてのカテゴリを前記文書画像に対応づけて記憶することを特徴とする請求項1に記載の文書処理装置。

【請求項7】　文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリを前記画像特徴と対応づけて記憶しておく文字カテゴリ記憶手段と、単語とその単語の各文字を前記カテゴリに置き換えたカテゴリ単語とを対応づけて記憶するカテゴリ単語辞書と、入力された文書画像中の文字ごとに画像を切り出すテキスト領域抽出手段と、該テキスト領域抽出手段によって切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ記憶手段に記憶されているカテゴリに分類する擬似文字認識手段と、該擬似文字認識手段によってカテゴリに分類されたカテゴリの列であるカテゴリ列を前記カテゴリ単語辞書から検索するカテゴリ単語検索手段を具備することを特徴とする文書処理装置。

【請求項8】　さらに、前記テキスト領域抽出手段によって切り出された各文字画像を該文字画像の外接矩形の大きさおよびその位置のいずれか1つ以上を用いて句読点か否かを判断する句読点検出手段を具備し、前記カテゴリ単語検索手段は、前記句読点検出手段によって句読点と判断された文字画像間の文字画像に対応する前記類似文字認識手段によって分類されたカテゴリ列を検索単位とすることを特徴とする請求項7に記載の文書処理装置。

【請求項9】　さらに、前記カテゴリ単語辞書に記憶されているカテゴリ単語の品詞およびそのカテゴリ単語に対応する単語間の接続関係を記憶する品詞接続辞書を具備し、前記カテゴリ単語検索手段は、前記品詞接続辞書に記憶されているカテゴリ単語の品詞および該カテゴリ単語に対応する単語間の接続関係に基づいてカテゴリ列を前記カテゴリ単語辞書から検索することを特徴とする請求項7に記載の文書処理装置。

【請求項10】　さらに、前記カテゴリ単語検索手段により検索されたカテゴリ列に対応する単語が複数存在する場合に、該カテゴリ列に対応する文字画像に対して文字認識を行なう文字認識手段を具備することを特徴とする請求項7に記載の文書処理装置。

【請求項11】　前記擬似文字認識手段は、文字の画像特徴とカテゴリを代表する画像特徴が閾値内の距離に存在する複数のカテゴリに分類し、前記カテゴリ単語検索手段は、前記擬似文字認識手段によって分類された複数のカテゴリの列であり、その複数のカテゴリの組合せであるカテゴリ列を前記カテゴリ単語辞書から検索することを特徴とする請求項7に記載の文書処理装置。

【請求項12】　文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリをその画像特徴と対応づけて記憶しておく文字カテゴリ記憶手段を具備した文書処理装置における文書処理方法において、入力された文書画像中の文字ごとに画像を切り出し、切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ記憶手段に記憶されているカテゴリに分類し、分類された前記各文字画像のカテゴリを前記入力された文書画像と対応

づけて記憶し、検索時に入力された検索式中のキーワードの各文字を前記文字カテゴリ記憶手段に記憶されている対応するカテゴリに変換し、カテゴリに変換された検索式を満たすカテゴリを有する文書画像を取り出すことを特徴とする文書処理方法。

【請求項13】　文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリを該画像特徴と対応づけて記憶しておく文字カテゴリ記憶手段と、単語とその単語の各文字を前記カテゴリに置き換えたカテゴリ単語とを対応づけて記憶するカテゴリ単語辞書を具備した文書処理装置における文書処理方法において、入力された文書画像中の文字ごとに画像を切り出し、切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ記憶手段に記憶されているカテゴリに分類し、カテゴリに分類されたカテゴリの列であるカテゴリ列を前記カテゴリ単語辞書から検索することを特徴とする文書処理方法。

【請求項14】　コンピュータに実行させるプログラムおよび辞書を読取可能に記憶した記憶媒体において、前記辞書は、文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリを前記画像特徴と対応づけて記憶しておく文字カテゴリ辞書であり、前記プログラムは、入力された文書画像中の文字ごとに画像を切り出すテキスト領域抽出手段と、該テキスト領域抽出手段によって切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ辞書に記憶されているカテゴリに分類する擬似文字認識処理と、該擬似文字認識処理によって分類された前記各文字画像のカテゴリを前記入力された文書画像と対応づけて記憶しておく擬似文字認識結果記憶処理と、検索時に入力された検索式中のキーワードの各文字を前記文字カテゴリ辞書に記憶されている対応するカテゴリに変換するキーワード変換処理と、該キーワード変換処理によってカテゴリに変換された検索式を満たすカテゴリを有する文書画像を前記擬似文字認識結果記憶処理によって記憶されているものから取り出す文書検索処理を前記コンピュータに実行させることを特徴とする記憶媒体。

【請求項15】　コンピュータに実行させるプログラムおよび辞書を読取可能に記憶した記憶媒体において、前記辞書は、文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリを前記画像特徴と対応づけて記憶しておく文字カテゴリ辞書と、単語とその単語の各文字を前記カテゴリに置き換えたカテゴリ単語とを対応づけて記憶するカテゴリ単語辞書であり、前記プログラムは、入力された文書画像中の文字ごとに画像を切り出すテキスト領域抽出処理と、該テキスト領域抽出処理によって切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ辞書に記憶されているカテゴリに分類する擬似文字認識処理と、該擬似文字認識処理によってカテゴリに分類されたカテゴリの列であるカテゴリ列を前記カテゴリ単語辞書から検索するカテゴリ単語検索処理を前記コンピュータに実行させることを特徴とする記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、文書を画像として入力して蓄積する文書処理装置に関するものであり、特に、文書画像中のテキスト内容を検索する検索機能を有する文書処理装置に関するものである。

【0002】

【従来の技術】文書をイメージスキャナ等の画像入力装置で画像に変換して電子的に蓄積し、後から検索することを可能とする文書ファイリング装置が実用化されている。しかしながら、その多くは入力した画像1枚ごとにキーワード等の検索のための属性を人手で付与しなければならず、非常に労力を要していた。

【0003】本来、文書の検索ではテキスト内容によるフルテキスト検索が望ましい。しかし、これはＤＴＰ等によって作成された電子文書に対しては可能であるが、文書画像に対しては直接行なうことはできない。このため、例えば、特開昭62－44878号公報では、文書中のテキスト部分に対して文字認識を行ない、コード化されたテキスト内容でフルテキスト検索を可能にしている。しかしながら、文字認識、特に多くの文字種を持つ日本語などにおいては、一般的に数百次元の特徴量ベクトルを求め、約3，000文字種以上の文字種の特徴量との照合を行なうため、特徴ベクトルの照合処理に非常に多大な計算機パワーが必要であった。また、文字認識率も高くないため、検索すべきキーワードが誤認されてしまう可能性があるという問題点があった。さらに特開昭62－44878号公報では、文字認識処理中に得られた各文字の候補を保持しておき、誤認による検索のもれを減少させている。また、特開昭62－285189号公報では、文字を認識後、形態素解析を利用して日本語として妥当な文字列を得ることで、誤認識した文字の修正を自動的に行なっている。特開平5－54197号公報では、誤認識された文字を修正するために、漢字を複数の代表文字によって置き換え、取り扱う字種を減らして確率遷移行列を利用して単語を同定している。しかし、これらの文献に記載されている技術は、基本的には文字認識処理を行なうために、文書登録時に多大な計算機パワーを要し、最終的に得たいものが検索時に指定した単語を含む文書画像であるとするならば、文字認識された結果はほとんどが無駄なものとなってしまう。

【0004】田中他，「日本語文書画像に対する文字列検索機能の実現」，情報処理学会情報メディア研究会資料19－1，1995年1月では、各文字画像から得られる特徴量を取り出して文字認識するのではなく、特徴量をそのまま36ｂｉｔのコードに変換する。次に検索キーワード画像の特徴量を抽出して特徴量のマッチングによって文字列検索を実現している。しかし、検索キー

ワードを画像として入力するか、あるいは文字フォントイメージによって画像を生成する必要があり、フォントの変動には弱いという欠点がある。

【0005】Reynar, J. et al, "Document Reconstruction:A Thousand Words from One Picture", in Proc. of 4th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, April 1995には、ヨーロッパ系言語（英語）のテキスト画像中の文字をその大きさ、位置によって少数のカテゴリに分類し、その並びによって単語として識別しようとする試みが開示されている。また、米国特許第5325444号明細書（1994）あるいは米国特許第5438630号明細書（1995）には、"Word Shape"などと呼ばれる単語単位での画像的な特徴を用いて、OCRを用いずに特定の単語の出現頻度を計測したり単語を同定する技術が開示されている。しかしながら、日本語や中国語などの多くの文字種を含む言語に対して、手がかりとするような特徴を直感的に設定することは困難である。また、ヨーロッパ系の言語と異なり、単語間のスペースが存在しないので単語単位で画像中から直接得ることができない。このため、直接的には開示されている手法を用いて日本語などのテキストを単語で識別することは困難であった。

【0006】また、特開平4－199467号公報には、誤認識しやすい文字種同士をグループ化し、グループに対して文字コードを割り当てておき、検索時にもグループを示す文字コードを用いて検索を行なうことが記載されている。この文献の方法では、一度文字認識処理を行なって文字コードを得た後、その文字コードをグループを示す文字コードへ変換している。そのため、グループ化によって検索漏れは防げるものの、文字認識のための多大な計算機パワーおよび時間が必要であることには変わりはない。

【0007】また、特開平7－152774号公報には、検索条件式の検索文字列を、誤認識しやすい文字について複数の候補により展開し、検索を行なうことが記載されている。さらに特開平6－103319号公報には、正常に変換できない文字が存在するとき、その文字をあいまいなまま残しておき、あいまいなデータを対象に検索を行なうことが記載されている。これらの文献に記載されている技術によれば、いずれも検索漏れを減少させることができる。しかし、これらの文献に記載されている技術においても、文字認識を行なうための多大な計算機パワーおよび時間が必要となる。

【0008】

【発明が解決しようとする課題】本発明は、上述した事情に鑑みてなされたもので、文書登録時に少ない計算機パワーでしかも高速に登録処理がおこなえるとともに、検索時には漏れの少ない検索を実現することのできる文書ファイリングを提供することを目的とするものである。

【0009】

【課題を解決するための手段】請求項1に記載の発明は、文書処理装置において、文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリを前記画像特徴と対応づけて記憶しておく文字カテゴリ記憶手段と、入力された文書画像中の文字ごとに画像を切り出すテキスト領域抽出手段と、該テキスト領域抽出手段によって切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ記憶手段に記憶されているカテゴリに分類する擬似文字認識手段と、該擬似文字認識手段によって分類された前記各文字画像のカテゴリを前記入力された文書画像と対応づけて記憶しておく擬似文字認識結果記憶手段と、検索時に入力された検索式中のキーワードの各文字を前記文字カテゴリ記憶手段に記憶されている対応するカテゴリに変換するキーワード変換手段と、該キーワード変換手段によってカテゴリに変換された検索式を満たすカテゴリを有する文書画像を前記擬似文字認識結果記憶手段から取り出す文書検索手段を具備することを特徴とするものである。

【0010】請求項2に記載の発明は、請求項1に記載の文書処理装置において、前記文字カテゴリ記憶手段に記憶されているカテゴリは、文字画像の特徴ベクトルによるクラスタリングによって分類を行なったものであることを特徴とするものである。

【0011】請求項3に記載の発明は、請求項1に記載の文書処理装置において、前記擬似文字認識結果記憶手段内の文書画像に対応づけて記憶しているカテゴリは、文書画像内において隣り合う2つの文字画像のカテゴリをキーとして該キーが出現する文書の識別子を記憶するバイグラムテーブルとして記憶されており、前記文書検索手段は、前記キーワード変換手段によって変換されたカテゴリを前記バイグラムテーブルから検索することを特徴とするものである。

【0012】請求項4に記載の発明は、請求項1に記載の文書処理装置において、前記文字カテゴリ記憶手段は、1つの文字を複数のカテゴリに記憶している場合があり、前記キーワード変換手段は、1つの検索キーワードに対して前記文字カテゴリ記憶手段内に記憶されているすべてのカテゴリに変換することを特徴とするものである。

【0013】請求項5に記載の発明は、請求項1に記載の文書処理装置において、前記文字カテゴリ記憶手段は、1つの文字を複数のカテゴリに記憶している場合があるとともにそれぞれのカテゴリに分類される確率を記憶し、前記文書検索手段は、前記文字カテゴリ記憶手段内の確率に応じて文書画像を前記擬似文字認識結果記憶

手段から取り出すことを特徴とするものである。

【００１４】請求項６に記載の発明は、請求項１に記載の文書処理装置において、前記テキスト領域抽出手段は、複数の文字切り出し解釈が存在する場合は該解釈すべてについて切り出しを行ない、前記擬似文字認識手段は、前記テキスト領域抽出手段により切り出されたすべての切り出し結果に対してカテゴリに分類し、擬似文字認識結果記憶手段は、前記擬似文字認識手段により分類されたすべてのカテゴリを前記文書画像に対応づけて記憶することを特徴とするものである。

【００１５】請求項７に記載の発明は、文書処理装置において、文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリを前記画像特徴と対応づけて記憶しておく文字カテゴリ記憶手段と、単語とその単語の各文字を前記カテゴリに置き換えたカテゴリ単語とを対応づけて記憶するカテゴリ単語辞書と、入力された文書画像中の文字ごとに画像を切り出すテキスト領域抽出手段と、該テキスト領域抽出手段によって切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ記憶手段に記憶されているカテゴリに分類する擬似文字認識手段と、該擬似文字認識手段によってカテゴリに分類されたカテゴリの列であるカテゴリ列を前記カテゴリ単語辞書から検索するカテゴリ単語検索手段を具備することを特徴とするものである。

【００１６】請求項８に記載の発明は、請求項７に記載の文書処理装置において、さらに、前記テキスト領域抽出手段によって切り出された各文字画像を該文字画像の外接矩形の大きさおよびその位置のいずれか１つ以上を用いて句読点か否かを判断する句読点検出手段を具備し、前記カテゴリ単語検索手段は、前記句読点検出手段によって句読点と判断された文字画像間の文字画像に対応する前記類似文字認識手段によって分類されたカテゴリ列を検索単位とすることを特徴とするものである。

【００１７】請求項９に記載の発明は、請求項７に記載の文書処理装置において、さらに、前記カテゴリ単語辞書に記憶されているカテゴリ単語の品詞およびそのカテゴリ単語に対応する単語間の接続関係を記憶する品詞接続辞書を具備し、前記カテゴリ単語検索手段は、前記品詞接続辞書に記憶されているカテゴリ単語の品詞および該カテゴリ単語に対応する単語間の接続関係に基づいてカテゴリ列を前記カテゴリ単語辞書から検索することを特徴とするものである。

【００１８】請求項１０に記載の発明は、請求項７に記載の文書処理装置において、さらに、前記カテゴリ単語検索手段により検索されたカテゴリ列に対応する単語が複数存在する場合に、該カテゴリ列に対応する文字画像に対して文字認識を行なう文字認識手段を具備することを特徴とするものである。

【００１９】請求項１１に記載の発明は、請求項７に記載の文書処理装置において、前記擬似文字認識手段は、文字の画像特徴とカテゴリを代表する画像特徴が閾値内の距離に存在する複数のカテゴリに分類する擬似文字認識手段と、前記カテゴリ単語検索手段は、前記擬似文字認識手段によって分類された複数のカテゴリの列であり、その複数のカテゴリの組合せであるカテゴリ列を前記カテゴリ単語辞書から検索することを特徴とするものである。

【００２０】請求項１２に記載の発明は、文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリをその画像特徴と対応づけて記憶しておく文字カテゴリ記憶手段を具備した文書処理装置における文書処理方法において、入力された文書画像中の文字ごとに画像を切り出し、切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ記憶手段に記憶されているカテゴリに分類し、分類された前記各文字画像のカテゴリを前記入力された文書画像と対応づけて記憶し、検索時に入力された検索式中のキーワードの各文字を前記文字カテゴリ記憶手段に記憶されている対応するカテゴリに変換し、カテゴリに変換された検索式を満たすカテゴリを有する文書画像を取り出すことを特徴とするものである。

【００２１】請求項１３に記載の発明は、文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリを該画像特徴と対応づけて記憶しておく文字カテゴリ記憶手段と、単語とその単語の各文字を前記カテゴリに置き換えたカテゴリ単語とを対応づけて記憶するカテゴリ単語辞書を具備した文書処理装置における文書処理方法において、入力された文書画像中の文字ごとに画像を切り出し、切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ記憶手段に記憶されているカテゴリに分類し、カテゴリに分類されたカテゴリの列であるカテゴリ列を前記カテゴリ単語辞書から検索することを特徴とするものである。

【００２２】請求項１４に記載の発明は、コンピュータに実行させるプログラムおよび辞書を読取可能に記憶した記憶媒体において、前記辞書は、文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリを前記画像特徴と対応づけて記憶しておく文字カテゴリ辞書であり、前記プログラムは、入力された文書画像中の文字ごとに画像を切り出すテキスト領域抽出手段と、該テキスト領域抽出手段によって切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ辞書に記憶されているカテゴリに分類する擬似文字認識処理と、該擬似文字認識処理によって分類された前記各文字画像のカテゴリを前記入力された文書画像と対応づけて記憶しておく擬似文字認識結果記憶処理と、検索時に入力された検索式中のキーワードの各文字を前記文字カテゴリ辞書に記憶されている対応するカテゴリに変換するキーワード変換処理と、該キーワード変換処理によってカテゴリに変換された検索式を満たすカテゴリを有する文書画像を前記擬似文字認識結果記憶処理によって記憶されているも

のから取り出す文書検索処理を前記コンピュータに実行させることを特徴とするものである。

【００２３】請求項１５に記載の発明は、コンピュータに実行させるプログラムおよび辞書を読取可能に記憶した記憶媒体において、前記辞書は、文字の画像特徴をもとに類似した文字ごとに分類されたカテゴリを前記画像特徴と対応づけて記憶しておく文字カテゴリ辞書と、単語とその単語の各文字を前記カテゴリに置き換えたカテゴリ単語とを対応づけて記憶するカテゴリ単語辞書であり、前記プログラムは、入力された文書画像中の文字ごとに画像を切り出すテキスト領域抽出処理と、該テキスト領域抽出処理によって切り出された各文字画像を所定の画像特徴をもとに前記文字カテゴリ辞書に記憶されているカテゴリに分類する擬似文字認識処理と、該擬似文字認識処理によってカテゴリに分類されたカテゴリの列であるカテゴリ列を前記カテゴリ単語辞書から検索するカテゴリ単語検索処理を前記コンピュータに実行させることを特徴とするものである。

【００２４】

【発明の実施の形態】図１は、本発明の文書処理装置の第１の実施の形態を示す構成図である。図中、１はプロセッサ、２は表示装置、３はキーボード、４はマウス、５はスキャナ、６はプリンタ、７は外部記憶装置、１１は類似文字分類部、１２は擬似文字認識部、１３は検索実行部である。プロセッサ１には、操作を指示するためのキーボード３、マウス４、結果を表示するためのディスプレイ２、文書を入力するためのイメージスキャナ５、結果を印字出力するプリンタ６、プログラムや処理のためのデータを保持する外部記憶装置７等が接続されている。プロセッサ１は、実際の処理を行なう部分であり、実際の処理は外部記憶装置７に蓄えられたソフトウエアによって実行される。プロセッサ１は、例えば通常のコンピュータ本体等で構成される。外部記憶装置７としては、例えば高速アクセスが可能なハードディスク等で構成することができる。外部記憶装置７は、文書画像を大量に保持するために光ディスクなどの大容量デバイスを用いるような構成をとっても構わない。

【００２５】プロセッサ１で行なわれる処理は、類似文字分類部１１、擬似文字認識部１２、検索実行部１３の３つで構成される。類似文字分類部１１は、対象となる文字を、画像特徴を基にして類似文字から構成されるカテゴリに分類する。ここでは文書の登録の際に必要となる類似文字カテゴリテーブル、および検索の際に必要となる文字コード・カテゴリ対応テーブルを作成する。実際の文書の登録および検索にはこれらの２つのテーブルがあればよいので、ここでの処理は文書画像の入力に先だって行なわれるのみである。類似文字カテゴリテーブルは、カテゴリを代表する文字の文字コード、実際にそれに属する複数の文字の文字コード、そのカテゴリを代表する画像特徴ベクトルを対にして記憶しているものである。文字コード・カテゴリ対応テーブルは、類似文字カテゴリテーブルの逆引きテーブルであり、検索キーワードを代表文字コード列に変換するために用いられる。

【００２６】擬似文字認識部１２は、入力された文書画像からテキスト領域を抽出し、各領域内に含まれるそれぞれの文字を類似文字カテゴリに分類して、その代表文字コードを割り当て、これらを対応する文字の画像上の位置とともに文書画像を外部記憶装置７に記憶する。

【００２７】検索実行部１３は、利用者に検索式の入力を促し、入力がなされたならばその検索式に含まれるキーワードを文字コード・カテゴリ対応テーブルによってカテゴリの代表文字コード列に変換し、その変換されたキーワードのコード列を含む文書画像を取り出し、見つかったキーワードの位置とともに利用者に提示する。

【００２８】以下、それぞれの処理の詳細について説明する。図２は、類似文字分類部の処理の一例を示すフローチャートである。類似文字分類部１１は、各類似文字カテゴリに含まれる文字画像のトレーニングサンプルを入力として、類似文字カテゴリテーブルおよび文字コード・カテゴリ対応テーブルを作成する。トレーニングサンプルは二値の文字画像とそれに対応する文字コードから構成され、さまざまなフォント、二値化のしきい値の異なるものなどをすべての文字種について用意する。

【００２９】まず、Ｓ２１において、前処理として各文字画像の大きさの正規化を行なう。ここでは正規化された大きさを６４×６４（画素）としておく。次に特徴抽出を行なう。ここではペリフェラル特徴を用いている。図３は、ペリフェラル特徴の説明図である。ペリフェラル特徴は、図３に示すように、文字の外接矩形のそれぞれの辺から走査し、白画素から黒画素に変化する点までの距離を特徴とするもので、最初に変化する位置と２度目に変化する位置を取り出す。ここでは、水平および垂直方向にそれぞれ８つの領域に分割して走査することとし、８×４×２の合計６４次元の特徴ベクトルを取り出す。図３では、外接矩形の左辺から走査した場合を示しており、最初に白画素から黒画素に変化する点までの走査軌跡を破線の矢印で示している。通常の文字認識ではさらに他の特徴量も併用して識別精度を向上させることを行なっているが、ここでは少数の類似文字カテゴリに分類するだけでよいので、少ない次元数の特徴ベクトルで十分の精度が期待できる。なお、ペリフェラル特徴に代えて、あるいはペリフェラル特徴とともに、他の特徴を抽出して特徴ベクトルを形成してもよい。

【００３０】トレーニングサンプルの各文字について特徴ベクトルが得られたならば、Ｓ２２において、同一の文字種、すなわち「亜」ならば同じ「亜」であって異なるフォントや二値化の異なるものなどについて特徴ベクトルの平均をとり、各文字種ごとの代表ベクトルを作成する。この代表ベクトル間の距離が特徴空間内で近いものが類似文字である。Ｓ２３において、この代表ベクト

ルが近くに集まっているものをグループとしてまとめるクラスタリング処理を行なう。クラスタリングは、例えば、Duda, Hart著, “Pattern Classification and Scene Analysis”, Wiley-Interscience社刊に記載されている方法などを用いることができる。この方法はまず、初めに階層的クラスタリングを施し、これを最初のクラスタの仮定としてクラスタごとの重心と各特徴ベクトルとの自乗誤差の総和が最小になるように最適化を行なうものである。

【0031】図4は、階層的クラスタリングの処理の一例を示すフローチャートである。まずS31において、所望のクラスタ数をm、文字種の総数をn、初期クラスタを$X=\{c_i \mid i=1,\ldots,n\}$とし、$c_i$は類似している文字種の代表特徴ベクトルが保持される。c_iの初期値として、各文字種の代表特徴ベクトルを1つずつ入れる。S32において、現在のクラスタ数と所望のクラスタ数mとを比較し、もし現在のクラスタの数がmに等しければ、その時点のXをクラスタリングの結果として処理を終わる。そうでない場合はS33へ進む。S33において、特徴空間におけるクラスタの距離dが最も小さい2つのクラスタの組を見つけ出し、これを一つのクラスタに統合する。そしてS32へ戻る。

【0032】所望のクラスタ数mは任意に与えることができるが、ここでは仮に500に設定しておく。JIS第一水準では約3，000字種が存在するため、1クラスタ当たり平均6字種が含まれることになる。この処理の中で、クラスタ間の距離dを計算する方法としては種々のものが考えられる。ここでは、2つのクラスタ内の特徴ベクトルを1つずつ取り出して組を作り、その中で最も近い位置にあるベクトルの組の距離を2つのクラスタの距離とする方法を用いることにする。

【0033】この階層的クラスタリングの結果は最適なクラスタリングとはいえないため、これを出発点として、図2のS24においてクラスタの最適化を行なう。最適化は各クラスタ内の特徴ベクトルの平均値と各特徴ベクトルとの距離の二乗和をとり、すべてのクラスタについての総和を判定関数とする。この判定関数の値が小さいほどクラスタ内の特徴ベクトルが密集しており、より良いクラスタリングであるといえる。これを最小とするようなクラスタリングを見つけることは一般的には困難であるが、擬似的に最適化を施すことが可能である。

【0034】図5は、クラスタリングの最適化処理の一例を示すフローチャートである。まずS41において、任意の特徴ベクトルxを取り出す。そしてS42において、特徴ベクトルxが現在属しているクラスタをc_iとして、そこに登録されている特徴ベクトルがxのみであるか否かを判定し、特徴ベクトルxのみである場合はS41へ戻る。そうでない場合は、すべてのクラスタc_jに対して以下の計算を行なう。

$$(j \neq i \text{の時}) \quad a = n_j / (n_j + 1) \| x - m_j \|^2$$

$$(j = i \text{の時}) \quad a = n_i / (n_i - 1) \| x - m_i \|^2$$

ただし、n_jはc_jに登録されている特徴ベクトルの個数、m_jはc_jに属する特徴ベクトルの平均である。上記の式は特徴ベクトルxをc_jに移動させた時の判定関数の変化量を示している。

【0035】S44において、S43で計算されたaの値が最小となるjがi以外であるか否かを判定し、aの値が最小となるjがi以外である場合はS45において特徴ベクトルxをクラスタc_jへ移動させる。

【0036】S46において、すべての特徴ベクトルについてクラスタの移動ができなくなったか否かを判定し、まだ移動が可能な場合には、S41へ戻って次の特徴ベクトルをxとしてS42以下の処理を繰り返す。すべての特徴ベクトルについてクラスタの移動ができなくなった場合は、その時点でのクラスタを結果とし、処理を終了する。

【0037】このようにして類似文字のクラスタリングが行なわれる。この図5に示した処理において、S41で任意の文字を取り出す際の方法をさまざまに変えて同様の処理を施し、評価関数（各クラスタ内の特徴ベクトルの平均値と各特徴ベクトルとの距離の二乗和の総和）を最小とするものを結果として採用する。

【0038】図2に戻り、S25において、それぞれのクラスタに基づき、類似文字カテゴリテーブルを作成して記憶する。この類似文字カテゴリテーブルは、文書の登録の際に用いられる。図6は、類似文字カテゴリテーブルの一例の説明図である。図6に一部示した類似文字カテゴリテーブルは、各カテゴリごとに、属する文字の文字コード、カテゴリ特徴の代表ベクトル、およびカテゴリを代表する文字コードから構成されている。カテゴリ特徴ベクトルは属する文字の特徴ベクトルの平均である。カテゴリを代表する文字コードはそのカテゴリに属する文字の文字コードのうち、任意の1つが当てられる。図6では、文字コードの代わりに文字自体を記載している。

【0039】さらに、S26において、検索処理で検索キーワードを代表文字コード列に変換するために、類似文字カテゴリテーブルの逆引きテーブルとして文字コード・カテゴリ対応テーブルを同時に作成する。図7は、文字コード・カテゴリ対応テーブルの一例の説明図である。文字コード・カテゴリ対応テーブルは、図7に示すように、文字コードと、その文字コードに対応するカテゴリの代表文字コードを組にして作成する。

【0040】次に、擬似文字認識部12において行なわれる文書の登録処理について述べる。図8は、擬似文字認識部の処理の一例を示すフローチャートである。まず、利用者は接続されているイメージスキャナ5などに

よって登録したい文書を画像として入力する。あるいは、ＦＡＸやネットワークなどで伝送されて入力される場合もある。ここではモノクロ二値画像を入力と想定しているが、グレースケールあるいはカラー文書として入力し、擬似文字認識処理に対しての入力の際に、しきい値処理によって二値画像に変換してもよい。入力された二値画像に対して、まず前処理としてノイズ除去、スキュー補正などが行なわれる。

【００４１】Ｓ５１において、二値画像の中に含まれる文字領域が抽出される。この処理は例えば、秋山，増田，「周辺分布、線密度、外接矩形特徴を併用した文書画像の領域分割」；電子情報通信学会論文誌Ｄ－ＩＩ，Ｖｏｌ．Ｊ６９，Ｎｏ．８などに開示されている周辺分布による領域分割手法などを用いることができる。もちろん、領域分割処理方法としては多くの手法が提案されており、ここで述べる周辺分布に基づく手法に限ったものではないことはいうまでもない。図と判定された部分は処理対象から除かれる。分割された文字ブロック領域は矩形領域として順にブロックＩＤと呼ばれる番号が付与され、メモリ内に保持される。

【００４２】図９は、文字領域抽出結果の一例を示す説明図である。図９（Ａ）は入力された文書画像の一例を示しており、ハッチングを施した部分が文字が並んだ行を示しており、×を付した部分が図の領域である。例えば、このような二値の文書画像が入力されると、図９（Ｂ）に太枠で示すような各文字ブロック領域と図表領域に分割され、文字ブロック領域に対してブロックＩＤが付与される。図９（Ｂ）においてはブロックＩＤ１～６が付与されている。

【００４３】図８に戻り、Ｓ５２において、文字領域はさらに行ごとに分割され、さらに文字ごとに分割される。この文字の切り出し処理についても種々の手法が提案されており、いずれの手法を用いてもよい。

【００４４】Ｓ５３において、切り出された各文字画像ごとに類似文字カテゴリの代表文字コードへ変換する。図１０は、代表文字コード列への変換処理の一例を示すフローチャートである。まず、明らかに検索キーワードになりえない句読点を取り出しておく。Ｓ６１において、文字画像が句読点であるか否かを判定する。句読点の判定は、その文字画像の外接矩形の幅、高さがしきい値Ｔｗ，Ｔｈ以下であるもので、上端が文字行の中心より下にあって、右に隣接する文字までの距離がしきい値Ｔｒより大きいという条件を満たすものである。しきい値Ｔｗ，Ｔｈ，Ｔｒは日本語文字幅と高さがほぼ同一であるという条件から、文字行の高さをｈとすると、例えば、Ｔｗ＝Ｔｈ＝Ｔｒ＝ｈ／２と設定すればよい。句読点と判定された文字については、Ｓ６２において、文字カテゴリとして句読点を示す“。”を割り当てる。

【００４５】句読点でない場合、類似文字分類処理と全く同様に、Ｓ６３において大きさの正規化がなされ、画像特徴が計算される。ここでは、類似文字分類処理時にペリフェラル特徴を抽出したので、それに合わせてペリフェラル特徴を計算する。次にＳ６４において、この未知文字の特徴ベクトルがどの類似文字カテゴリに属するかを判定する。すなわち、未知文字の特徴ベクトルと類似文字カテゴリの代表ベクトルとのユークリッド距離を計算して比較する。代表ベクトルは、類似文字カテゴリテーブルに登録されているので、これを用いることができる。Ｓ６５において、計算されたユークリッド距離が最も近いものをその文字カテゴリとして採用し、その代表文字コードを結果として出力する。ここでは簡単のために最短距離による識別方法を用いているが、この最短距離による識別方法以外にもさまざまな識別手法が考えられ、それらを用いることもできる。

【００４６】図１１は、代表文字コード列への変換処理の結果の一例を示す説明図である。いま、入力された文字画像が図１１（Ａ）に示すように「…文書画像解析…」であった場合に、まず最初の文字画像「文」を切り出し、特徴ベクトルを求める。次に類似文字カテゴリテーブルに記憶されている各カテゴリの代表ベクトルとの距離を求め、最短距離を持つカテゴリの代表文字コードを割り当てる。例えば、図６に示すような類似文字カテゴリテーブルが登録されているとき、順に文字画像すべてに対して代表文字コードへの変換を行なうと、この画像はカテゴリの代表文字コード列「…父家画倶絹肝…」に変換される。

【００４７】ここでは通常の文字認識は行なっておらず、少ない次元の特徴ベクトルを用いて少数の文字カテゴリとの照合を行なっているに過ぎない。類似文字カテゴリテーブルには類似文字コードが登録されてはいるが、文字認識を行なっていないのでこの類似文字コードはこの時点では使用されない。

【００４８】このように、代表文字コード列への変換処理は、少数の文字カテゴリとの照合ですむため、大幅な速度向上が実現できる。照合はユークリッド距離を用いており、計算量は特徴ベクトルの次元数と識別カテゴリの数にほぼ比例する。いま、識別する対象の文字種の数を３，０００、類似文字カテゴリの数を５００とし、特徴ベクトルの次元数を通常の文字認識の場合を３００、本手法の場合を６４とすると、トータルで照合のための計算量は１／２８以下に抑えることができる。日本語の文字認識の高速化手法として、少数次元の特徴ベクトルを用いて近い文字種を数十から数百取り出しておき（大分類）、さらに詳細な識別をさらに多次元の特徴ベクトルを用いて行なう（詳細分類）という階層的な識別手法が知られている。このような手法での大分類処理で本手法と同一次元数のベクトルを用いたと仮定しても、全文字種（３，０００）との照合が必要であり、さらに詳細分類が必要となるので、トータルの計算量は１／６以下になる。

【0049】図8に戻り、S53で得られた代表文字コード列をそのまま検索処理の時にサーチするのでは効率が悪いので、検索のためのインデックスを準備し、文書を登録するごとにその内容を更新する。ここではbi－gramによるインデックスを用い、S54においてbi－gramテーブルへの登録を行なう。bi－gramは文字列の中の2つの連続する文字からなる部分文字列を指す。すなわち、「父家画倶絹肝」という文字列の場合には、bi－gram「父家」、「家画」、「画倶」、「倶絹」、「絹肝」が得られる。これを代表文字コード列について取り出し、テーブルのインデックスにして、その文書画像IDとそのbi－gramの代表文字コード列内の位置（すなわち、何文字目）を保存しておく。

【0050】図12は、bi－gramテーブルの一例の説明図である。図12には、上述の例で用いた「文書画像解析」という文字列に対して得られた代表文字コード列「父家画倶絹肝」のbi－gramテーブルを示している。図12に示したbi－gramテーブルは2段階で構成されており、bi－gramをキーとしてその内容を示すテーブルへのポインタを格納する。ポインタによって示されるテーブルは、文書IDとその中のどの領域かを示すブロックIDと文字位置との組からなるテーブルとして構成され、入力された文書中の文字ブロック内に、対応するbi－gramが見つかるたびにそのエントリが追加されていく。bi－gramテーブルは公知の技術、例えば、bi－gramをキーとするB－treeまたはHashテーブルなどによって実現でき、高速な検索を可能とすることができる。なお、最初に句読点と判断されたものについてはbi－gramは生成されない。

【0051】図8に戻り、S55において、S53で得られた代表文字コード列を、文字ブロックごとにその画像上の位置とともに代表文字コードテーブルとして、入力画像とともに外部記憶装置7などに蓄える。図13は、代表文字コードテーブルの一例を示す説明図である。各代表文字コードと、その文字コードが画像上で占める矩形位置を対にして記憶している。図13では、代表文字コードの代わりに文字を記して示している。また、文字コードが画像上で占める矩形位置は、（左上x座標、左上y座標、幅、高さ）で表現している。以上の処理によって入力された文書画像についての登録処理が完了する。

【0052】最後に検索実行部13における検索処理について説明する。図14は、検索実行部の処理の一例を示すフローチャートである。検索実行部13は、利用者からの入力があるまで待っている。利用者がディスプレイ2を見ながら、例えばキーボード3で検索式を入力すると、検索実行部13はS71において入力された検索式を読み込む。検索式としては、種々の形態が可能であるが、ここでは、検索キーワードを論理和、論理積、論理否定などブール演算子で結合して構成されているものとする。

【0053】検索式を読み込むと、S72において検索式を解析して検索式内のキーワードを取り出し、S73において、検索式内のキーワードを文字コード・カテゴリ対応テーブルを参照してカテゴリの代表文字コード列に変換する。具体例として、検索式が「文書画像＊解析」である場合について考える。ここで、＊は論理積を表わす。この検索式は「文書画像」という単語と「解析」という単語を共に含む文書画像を検索せよという指示を意味する。2つのキーワードに対応する代表文字コード列は文字コード・カテゴリ対応テーブルを参照して、それぞれ「父家画倶」、「絹肝」に変換される。

【0054】次に、登録されている文書画像から得られた代表文字コード列の中に、この2つのキーワードから変換された代表文字コード列を含むものがあるか否かを調べ、あればその画像上の位置を記憶する。実際はS74においてキーワードに対応する代表文字コード列のbi－gramを作成し、これをS75において前述のbi－gramテーブルの中から検索し、対応する文書画像のIDとそのbi－gramの出現位置を得る。3文字以上の検索キーワードの場合は複数のbi－gramが生成され、それぞれのbi－gramが同一文書の同一文字ブロック中で連続して出現している必要がある。したがって、同一の文書画像IDとブロックIDについてそのbi－gramの出現位置を前から順にトレースし、連続していないものは結果から削除する。

【0055】上述の検索式の例では、キーワード「父家画倶」からbi－gram「父家」、「家画」、「画倶」が作成され、キーワード「絹肝」はそのままbi－gram「絹肝」となる。例えば、図12に示すようなbi－gramテーブルが登録されているとする。まずbi－gram「父家」が含まれる文書は、文書IDが00001、00015、00023の4つである。このうち、文書IDが00001の文書では、ブロックID1，2内の「父家」の位置のあとには「家画」というbi－gramが連続していることがわかる。しかし、文書IDが00015や00023の文書では、「家画」というbi－gramは連続していない。したがって、文書ID00001の文書が「父家画」という文字列を含むことが分かる。同様の処理を「画倶」についても調べて、最終的に「父家画倶」が含まれる文書の文書IDが得られる。「絹肝」は2文字単語なのでこのbi－gramテーブルを調べるだけでよい。こうして各検索キーワードが出現している文書画像IDとその出現位置が得られる。

【0056】最後にS76において検索式内の論理演算を施す。すなわち、各検索キーワードを含む文書画像IDの集合に対して論理演算を行ない、最終的に検索式に

合致する文書画像ＩＤの集合を得る。例えば、キーワードに対応する代表文字コード列「父家画倶」、「絹肝」を含む文書ＩＤの集合がそれぞれ（００００１，００００３１，００２０２）、（００００１，００００５４，００２０２）であった場合に、論理積を施すと、（００００１，００２０２）となる。すなわち、文書画像ＩＤ００００１の文書画像と、文書画像ＩＤ００２０２の文書画像が、代表文字コード列「父家画倶」、「絹肝」の両方を含んでいることになる。

【００５７】Ｓ７７において、このようにして得られた結果に含まれる文書画像ＩＤに対応する文書画像を例えば外部記憶装置７から取り出し、Ｓ７８においてディスプレイ２上に順に表示する。また、得られたブロックＩＤと文字位置をもとに、画像とともに記憶している画像上の代表文字コードテーブルから文字の位置が分かるので、対応する文字をハイライト表示する。ハイライト表示は白黒反転表示でもよいし、カラーディスプレイの場合は分かりやすい色を用いても構わない。結果を見て利用者が印刷指定をした場合は、文書画像をプリンタ６へ出力すればよい。

【００５８】次に、本発明の文書処理装置の第１の実施の形態における第１の変形例について説明する。この第１の変形例では、さらに検索の精度を上げるための改良について述べる。伊藤他，「階層的印刷漢字認識システムにおける字種を複数クラスタに登録する辞書構成法」，電子情報通信学会論文誌Ｄ－ＩＩ，Ｖｏｌ．Ｊ７８－Ｄ－ＩＩ，Ｎｏ．６，ｐｐ．８９６－９０５，１９９５年６月でも示されているように、同一字種の特徴ベクトルを平均した代表ベクトルを用いてクラスタリングを行なった場合には、実際の文字画像に対して正しく対応するカテゴリに識別できない場合が存在する。これを避けるために、上記の文献に開示されているε－ｃｏｍｐｏｎｅｎｔ拡張法を用いることができる。すなわち、文字種ごとの代表ベクトルを用いてクラスタリングした後、テストサンプルの文字画像それぞれの特徴ベクトルと各カテゴリの代表ベクトルとのユークリッド距離を調べ、最短のものおよびその最短距離にスカラーパラメータεを加えた距離以内に存在するすべてのカテゴリにその文字コードを類似文字として登録する。εの値は大きくなればなるほど擬似文字認識の精度は向上するが、カテゴリあたりに含まれる文字コードが増加するため、検索時に誤った結果を出力する可能性が増える。最適なεの値を決定するために、まずテストサンプルとは別の未知文字画像のセットを準備する。種々のεに対して拡張された類似文字カテゴリを用いて擬似文字認識処理を行ない、未知文字画像セットのすべての文字について識別されたカテゴリに正しくその文字コードが含まれるような最小の値にεをセットする。

【００５９】このようにした場合、検索のための文字コード・カテゴリ対応テーブルが１つの文字コードに対して複数の類似文字カテゴリが対応するようになる。図１５は、複数のカテゴリへの分類を許容した場合の文字コード・カテゴリ対応テーブルの一例の説明図である。図１５に示した例では、例えば文字「並」は、代表文字が「亜」であるカテゴリと、代表文字が「平」であるカテゴリの２つに分類されている。図１５では示されていないが、１つの文字が３つ以上のカテゴリに分類されることもある。

【００６０】このように１つの文字が複数のカテゴリに分類されているため、検索式中のキーワードを代表文字コード列に変換する際に、１つのキーワードに対して可能な代表文字コード列が複数得られることになる。例えば、文字コード・カテゴリ対応テーブルが図１５に示す内容であるとき、文字「文」と「像」はそれぞれ「父、交」と「倶、場」の２つのカテゴリに属している。この場合、上述の検索式の例で用いたキーワード「文書画像」は、４つの代表文字コード列「父家画倶」、「交家画倶」、「父家画場」、「交家画場」に変換される。これら４つの代表文字コード列の少なくとも１つを含む文書をすべて取り出し、これら４つのキーワードの論理和として内部的に処理すればよい。このような処理を行なうことによって、若干の処理時間が増えるが、漏れのない検索を行なうことができる。

【００６１】さらに１つの文字に対して複数のカテゴリが対応する場合、カテゴリの確からしさを合わせて保持しておくことで、内部的に展開された４つのキーワードの確からしさを示すことができる。例えば、文字「文」が「父」カテゴリに識別される確率が０．７、「交」カテゴリに分類される確率が０．３であり、「像」も同様に「倶」カテゴリに識別される確率が０．８、「場」カテゴリに識別される確率が０．２であるとする。この場合、「父家画倶」は０．７×０．８＝０．５６、「交家画倶」は０．３×０．８＝０．２４、「父家画場」は０．７×０．２＝０．１４、「交家画場」は０．３×０．２＝０．０６の確率で出現する。このように展開されたキーワードを確からしい順に並べかえることによって、検索された文書画像を確からしいものから順に利用者に提示することも可能となる。各文字が対応するカテゴリに分類される確からしさは、例えば、カテゴリの拡張時に用いた未知文字画像セットの同一文字種の文字がどれくらいの割合で対応するカテゴリに含まれたかを数え上げることで計算できる。

【００６２】次に、本発明の文書処理装置の第１の実施の形態における第２の変形例について説明する。これまでは、文字切り出しの段階での誤りがなく、各文字が確実に切り出されるものとしてきたが、現実には切り出し時の誤りも多く発生する。日本語文字だけで構成される場合は固定ピッチが想定できるが、英単語などが入ることが想定される場合は、横書きテキストの場合はへんとつくりに分離されることが往々にして起こる。もちろ

ん、読み取り時のかすれなどが原因で1つの文字が2つの文字に分かれたりすることも想定される。

【0063】いくつかの文字について可能な文字切り出し位置が複数存在する場合は、その可能な切り出し結果を保持した代表文字コード列を表現すればよい。このような場合を想定して以下のように代表文字コード列を表現することを考える。これは実施例1で述べた代表文字コードテーブルを次のように拡張することによって実現する。

【0064】図16は、複数の文字切り出し解釈が存在する場合の切り出し位置の具体例を示す説明図である。いま、文字の切り出し処理の対象とする画像が図16（A）に示されるような「文書印刷」であった場合を考える。「文」、「書」については文字間の間隙しか存在しないので、適切に文字を切り出すことができる。しかし、「印」の文字中に1か所、「刷」の中に2か所、垂直方向に白画素のみからなる切り出し位置候補が存在する。これら2文字の間も当然切り出し位置が存在するので、「印刷」からは図16（B）に示すように合計5つの部分文字（a1，a2，b1，b2，b3）が得られる。

【0065】これらについて、文字としての統合を試みる。統合は部分文字を左から順に見ていき、幅のしきい値を越えないものはすべて文字として見なすとする。幅のしきい値としては、例えば行の高さhを用いることができる。この例では、文字「文」と統合できるものはないので、そのまま1文字として登録する。「書」も同様である。文字「印」については、部分文字a1，a2を2つの文字として扱う場合と1つの文字として扱う場合の2つが可能な解釈がある。a2とb1を統合した場合は幅のしきい値を越えるため、統合はなされない。したがって、ここまでの2つの解釈を同じ文字画像領域に対して保持する必要がある。同様にb1以降を順に見ていくと、可能な解釈が（[b1]，[b2]，[b3]），（[b1b2]，[b3]），（[b1]，[b2b3]），（[b1b2b3]）の4通りある。ここで、[]は中の部分文字が1つの文字と見なされることを示している。

【0066】図17は、複数の文字切り出し解釈が存在する場合の切り出された文字列の関係の説明図である。上述のようにして文字としての統合を試みた際の可能な解釈の関係を図17に示している。図中の○は文字切り出しの解釈の区切りであり、□は1つの文字として扱う単位を示している。a1とa2については、上述のような2通り、b1～b3については4通りの解釈があるので、それらの各解釈にそって切り出した候補を並べて線で結んで示している。この例では全部で8通りの解釈が成り立つ。これらすべての解釈が保持される。

【0067】図18は、複数の切り出し解釈を許容した場合の代表文字コードテーブルの一例の説明図である。図17に示すような複数の解釈を表現するため、具体的には図18に示すように、代表文字コードテーブルを基本テーブルとサブテーブルに分割する。基本テーブルは図13に示した代表文字コードテーブルを拡張し、複数の文字切り出し解釈がある場合にその解釈を表現するサブテーブルへのポインタを、画像上の位置を示していたカラムに格納できるようにする。複数の解釈がある場合、図18では基本テーブルの代表文字コードに0をセットしている。サブテーブルは、ある切り出し位置から見て右に文字と見なされる部分文字領域とその画像上の位置、その後に接続するサブテーブルの番号によって構成されている。

【0068】図16に示された文字「印」について考えると、文字切り出し位置は部分文字a1の左とa2の左にある。サブテーブルは切り出し位置の左から順に番号が付与される。すなわち、a1の左を切り出し位置とした場合に、可能な文字としての解釈は[a1]と[a1a2]である。a1はa2の左の切り出し位置を共有しているので、[a1]に対してはサブテーブルの番号2が格納されている。[a1a2]のほうはこれ以上接続する文字はないので、0が格納されている。

【0069】次にa2の左の切り出し位置とした場合について、2番目のサブテーブルが作成される。この切り出し位置の右における文字としての解釈は[a2]しか存在しない。そのため、2番目のサブテーブルは[a2]のみが登録され、その後に接続するものがないので、次テーブル番号には0がセットされる。

【0070】文字「刷」についても同様に3つのサブテーブルが生成される。最初のサブテーブルは[b1]，[b1b2]，[b1b2b3]の3つの解釈が、2番目のサブテーブルは[b2]，[b2b3]という解釈が、3番目のサブテーブルには[b3]という解釈が生成される。当然、それぞれに切り出された文字について擬似文字認識処理が行なわれ、代表文字コードが割り当てられ、サブテーブルの代表文字コードの欄に格納される。図18ではそれぞれの切り出された文字に対する代表文字コードは{}で表現している。

【0071】図19は、複数の切り出し解釈を許容した場合の代表文字コードテーブルの作成処理の一例を示すフローチャートである。図18に示すような複数の切り出し解釈を許容した場合の代表文字コードテーブルを作成する際の処理の一例について説明する。まずS81において、初期値の設定を行なう。1行に含まれるk個の部分文字領域をp_1，p_2，・・・，p_kとし、そのリスト{p_1，p_2，・・・，p_k}を変数Lにセットする。このとき、k個の部分文字領域は、左から右にソートされているものとする。また、現在処理中の文字の切り出し解釈が複数存在するか否かを示すフラグSをFALSEに設定する。さらに、1文字として統合可能な部分文字領域のリストCを空にする。さらに、現在のサブ

テーブル番号を示す変数nを1に、統合途中の部分文字列の位置を示す変数mを1に、現在注目している部分文字領域の位置を示す変数iを1に、それぞれセットする。

【0072】S82において、現在注目している部分文字領域の位置が行末まで達したか否か、すなわちiとkを比較し、i≦kであればS83に進み、まだ処理されていない最左にある部分文字領域p_iを取り出し、リストCにp_iをセットする。S84において、その部分文字領域p_iあるいはその部分文字領域p_iを含む統合された部分文字領域と、その右に隣接する部分文字領域p_{m+1}との統合を考え、統合した場合の文字幅を計算する。S85において、計算された文字幅が閾値を越えたか否かを判定する。閾値を越えていない場合には、さらに統合することが可能であるので、S86においてフラグSをTRUEとし、リストCにp_{m+1}を追加し、変数mを1だけ増加させてS82へ戻る。この場合、変数iの値は変化せず、変数mの値が変化しただけであるので、S84においてさらに右に隣接する部分文字領域の統合が試みられることになる。このようにして、文字幅が閾値を越えるまで処理が繰り返される。S85において、統合した文字幅が閾値を越える場合には、S84において最後に試みられた統合は行なわず、S87へ進む。このとき、p_iからp_mまでは統合可能であることになる。それまでに統合可能な部分文字領域のリスト$\{p_i, \cdots, p_m\}$がリストCに格納されている。

【0073】S87において、リストCの要素がp_iのみであるか否かを判定する。すなわち、複数の部分文字領域が統合可能であるのか否かを判定する。複数存在する場合には、複数の部分文字領域について統合可能であるので、それらの部分文字領域からサブテーブルを作成する。S88において、リストCに格納されている部分文字領域の最左のものを含むすべての可能な統合文字領域を、部分文字領域の個数の少ない順に番号nのサブテーブルへ登録する。このとき、それぞれの統合文字領域について、大きさを正規化し、特徴量を計算して代表文字コードを割り当て、サブテーブルに登録する。また、次テーブル番号は、変数nの値に統合文字領域中の部分文字領域の個数を加えた値とし、サブテーブルの最後の統合文字領域の次テーブル番号は0にセットする。このようにしてi番目の部分文字領域から始まる統合文字領域について、サブテーブルが作成された。

【0074】S89において、次の部分文字領域から始まる統合文字領域についての処理を行なうべく、変数iを1だけ増加させ、注目する部分文字領域を次に移す。それとともに、リストCを空にリセットし、サブテーブルの番号を示す変数nを1だけ増加させ、変数mを変数iの値とする。そして、S82へ戻り、次の部分文字領域から部分文字領域の統合を試みる。

【0075】S87においてリストCの要素がp_iのみであった場合、さらにS90においてフラグSを調べる。フラグSがFALSEの場合、p_iは独立した文字である可能性のある部分文字領域である。S91において、その部分文字領域p_iの大きさを正規化し、特徴量を計算して代表文字コードを割り当て、基本テーブルに登録する。そして、次の部分文字領域について処理を行なうべく、変数iを1だけ増加させ、リストCを空にリセットする。そしてS82へ戻る。

【0076】S90においてフラグSがTRUEであった場合、部分文字領域p_iは、例えば、図16に示す例におけるa2やb3のように、統合可能な部分文字領域群の右端の部分文字領域である。この場合にはS93において、p_iを正規化して特徴量を計算し、代表文字コードを割り当ててn番目のサブテーブルを作成する。この時の次テーブル番号は0である。この部分文字領域p_iは右に統合する部分文字領域は存在しないので、基本テーブルの1つのエントリから連鎖するサブテーブルは終了する。そのため、S94においてサブテーブルの番号を示す変数nを1にリセットする。また、次の部分文字領域を処理すべく、変数iを1だけ増加させ、リストCを空にリセットし、フラグSをFALSEにリセットする。また、変数mをiにセットする。そして、S82へ戻り、新たに注目する部分文字領域からの処理を行なう。

【0077】行の右端の部分文字領域まで処理が終了すると、i＞kとなる。S82においてこの条件が判定されると、それ以上の統合処理は不要である。S95においてリストCが空か否かを判定し、空でない場合、リストCに残っている部分文字領域について、S87以降の処理を行ない、基本テーブルあるいはサブテーブルを作成する。リストCが空になると処理は終了する。このような処理によって、例えば、図18に示すような2層構造の代表文字コードテーブルが作成される。作成された代表文字コードテーブルは、入力された文書画像とともに登録される。

【0078】複数の文字切り出しの解釈を許容する場合、検索のためのインデックスであるbi-gramテーブルも複数の文字切り出しの解釈に対応できるように拡張を行なう。すなわち、bi-gramテーブルの2つの文字について、複数の文字切り出し解釈の1つであるか否か、そうであった場合に、どの文字切り出しの解釈に属するのかを明示する必要がある。そこで、bi-gramテーブルを以下のように拡張する。すなわち、図12に示した個々のbi-gramに対して格納されている文書画像上の位置のテーブルのうち、文書ID、ブロックIDは共通なのでそのままとし、第一文字、第二文字それぞれに対して、その位置を（p，n，m）の組で表わす。pはブロック内での文字位置、すなわち代表文字コードテーブル内での位置、nは文字切り出し解釈のサブテーブルの番号、mはサブテーブル内の位置を

それぞれ示す。

【0079】図20は、複数の切り出し解釈を許容した場合のｂｉ－ｇｒａｍテーブルの一例の説明図である。切り出しの解釈が一通りである場合は、nは0にセットされ、mは無視される。図20においてｂｉ－ｇｒａｍ「父家」の例はこれに該当する。

【0080】切り出しの解釈が複数あり、ｂｉ－ｇｒａｍの個々の文字がその中の1つである場合、nはサブテーブルの番号、mはそのサブテーブル内での位置を示す。図16に示した「印刷」の複数の文字切り出し候補の例で、例えば、「印」の文字が2つに分離されたｂｉ－ｇｒａｍ｛[ａ1]｝｛[ａ2]｝に対応する文字位置は、(116，1，1)，(116，2，1)となり、「印」「刷」が正確に切り出されたｂｉ－ｇｒａｍ｛[ａ1ａ2]｝｛[ｂ1ｂ2ｂ3]｝の文字位置は、(116，1，2)，(117，1，3)として格納される。このようにして、入力された文書画像の代表文字コード列から作成されたｂｉ－ｇｒａｍテーブルが登録され、検索の際に使用される。

【0081】また、検索の際には、入力された検索式の中のキーワードについて、文書画像の場合と同様にして代表文字コード列のｂｉ－ｇｒａｍを作成し、登録されているｂｉ－ｇｒａｍテーブルの中から検索すればよい。キーワードは、例えばキーボード3等によって入力されるので、検索実行部13は文字コードとして受け取るため切り出し位置による複数の解釈は存在せず、一意に決まる。文書画像から作成されたｂｉ－ｇｒａｍテーブルには、正しく切り出された場合のｂｉ－ｇｒａｍも登録されているので、検索の際にはそのようなｂｉ－ｇｒａｍとの一致が検出されることになる。

【0082】先に述べたように3文字以上のキーワードに対して同一文書に連続して存在しているか否かの判定する必要がある。いま2つのｂｉ－ｇｒａｍが連続して存在するか否かを判定する場合は、それらが同一文書ＩＤ、同一文字ブロックＩＤを持ち、前のｂｉ－ｇｒａｍの終わりの文字の位置を示す(ｐ，ｎ，ｍ)が、接続しているか否かを判定したいｂｉ－ｇｒａｍの始めの文字の位置と同一であればよい。このような場合に連続していると判定することができる。

【0083】なお、上述の第1の変形例で述べた複数のカテゴリへの分類を許容した場合の構成と、第2の変形例で述べた複数の文字切り出し解釈が存在する場合の構成を組み合わせて構成することも可能である。

【0084】次に、本発明の第2の実施の形態について説明する。上述のように、第1の実施の形態では、類似文字のカテゴリの列に変換して単純なマッチングによって検索を行なうので、文書中で単語として許容されないような文字列も検索してしまう可能性がある。この第2の実施の形態では、このような単語として許容されないような文字列を含む文書が検索されないようにし、さらに検索精度を向上させた例について説明する。

【0085】図21は、本発明の文書処理装置の第2の実施の形態を示す構成図である。図中、図1と同様の部分には同じ符号を付して説明を省略する。101は画像入力部、102は画像表示部、103は類似文字分類部、104はテキスト領域抽出部、105は擬似文字認識部、106はカテゴリ単語検出部、107はカテゴリ単語変換部、108は中央制御装置、109は記憶装置、111は文字カテゴリ保持部、112は擬似文字認識結果記憶部、113はカテゴリ単語辞書、114はコード変換テーブルである。

【0086】画像入力部101は、例えば図1に示したスキャナ5等で構成され、文書を画像として読み込む。画像表示部102は、例えば図1に示したディスプレイ2等で構成され、入力画像の表示や処理結果を確認するための表示などを行なう。類似文字分類部103は、図1における類似文字分類部11と同様のものであり、対象となる文字をその画像特徴をもとに類似文字からなるカテゴリに分類する。テキスト領域抽出部104は、図1における擬似文字認識部12の一部の機能を構成するものであり、文書画像中のテキスト領域を切り出し、さらに文字ごとに画像を切り出す。擬似文字認識部105は、図1における擬似文字認識部12の一部の機能を構成するものであり、各文字画像を類似する類似文字カテゴリに分類し、その代表文字コードを割り当てる。カテゴリ単語検出部106は、代表文字コード列から単語を構成する代表文字コード列を抽出する。カテゴリ単語変換部107は、カテゴリ単語を文字に変換する。中央制御装置108は、装置全体を制御する。

【0087】さらに記憶装置109は、図1に示す外部記憶装置7を含むものであり、中央制御装置108が装置全体を制御するためのプログラム等を格納するとともに、文字カテゴリ保持部111、擬似文字認識結果記憶部112、カテゴリ単語辞書113、コード変換テーブル114を含む。文字カテゴリ保持部111は、類似文字分類部103で分類されたカテゴリと対応する画像特徴を記憶する。例えば、上述の類似文字カテゴリテーブルや、文字コード・カテゴリ対応テーブルなどを記憶する。擬似文字認識結果記憶部112は、擬似文字認識部105で変換された代表文字コード列を保持する。カテゴリ単語辞書113は、少なくとも単語を構成する代表文字コード列と、品詞との対応関係を保持している。また、その代表文字コード列で表現される1以上の文字単語も保持する場合もある。さらに、品詞の接続関係を示す品詞接続辞書を保持する。コード変換テーブル114は、単語を表わす代表文字コード列と文字列との対応を記録している。カテゴリ単語辞書113にカテゴリ単語と対応づけて文字単語を保持している場合、コード変換テーブル114をカテゴリ単語辞書113で代用することも可能である。

【0088】以下、それぞれの処理の詳細について説明を行なう。まず、類似文字分類部103における処理は、上述の第1の実施の形態と同様であるので、ここでは説明を省略する。なお、類似文字分類部103で生成される類似文字カテゴリテーブルおよび文字コード・カテゴリ対応テーブルは、文字カテゴリ保持部111に保持される。類似文字分類部103は、解析する特徴量を決めてしまえば、処理ごとに行なう必要がなく、別の装置上で解析を行なって、その結果のみを文字カテゴリ保持部111に格納して使用することもできる。文字カテゴリ保持部111は、具体的には例えば図6に示すような類似文字カテゴリテーブル、および、例えば図7に示すような文字コード・カテゴリ対応テーブルを記憶する。

【0089】また、文字カテゴリ保持部111に記憶されている類似文字カテゴリテーブルおよび文字コード・カテゴリ対応テーブルを用いて、既存の単語辞書の文字コードを代表文字コード列で置き換えることによって、カテゴリ単語辞書113およびコード変換テーブル114を生成することができる。図22は、本発明の文書処理装置の第2の実施の形態におけるカテゴリ単語辞書の一例の説明図である。この例では、単語を示す代表文字コード列と、その代表文字コード列で示される単語の品詞と、その代表文字コード列で示される単語の文字列を対応づけている。既存の単語辞書には、文字単語と品詞とを対にして記憶しているものがあり、この文字単語に対応する代表文字コード列を得て、並べ替えることによって図22に示すようなカテゴリ単語辞書113が得られる。なお、図22に示すカテゴリ単語辞書113では、単語を示す代表文字コード列には、例えば活用変化する単語について、語幹となる単語だけでなく、語尾についても別に記憶している。そして、後述するように、品詞接続辞書を内蔵し、語幹と語尾との接続関係を示し、さらに接続される助動詞や助詞などを示すように構成している。あるいは、活用形をすべて記憶させておいてもよい。

【0090】図23は、本発明の文書処理装置の第2の実施の形態におけるカテゴリ単語辞書の別の例の説明図である。カテゴリ単語辞書113は、図22に示したようなが代表文字コード列と、文字コードによる単語、それと品詞の対応を示す表の形式のほかにも種々の形式で表現することができる。例えば、照合処理を効率的に行なうため、図23に示すような形式でカテゴリ単語辞書113を構成することができる。このカテゴリ単語辞書113は、例えば、青江，「トライとその応用」，情報処理，Vol．34，No．2，1993．2に紹介されているトライ（trie）を用い、各代表文字コードで始まるカテゴリ単語を全て保持するようにトライを構成している。そして、終端ノードまでたどることで、カテゴリ単語を抽出できる構造になっている。なお、図23では終端ノードを◎で示している。

【0091】図23に示した例では、例えば、「文字」、「文学」、「文学者」、「文学青年」、「文化」、「文化遺産」、「文化勲章」の7つの単語を照合できるカテゴリ単語辞書113を示している。7つの単語は、それぞれを代表文字コード列に直すと、「父手」、「父羊」、「父羊君」、「父羊君牛」、「父化」、「父化送屋」、「父化郵琴」となる。これらをそれぞれトライで表わすと、図23に示すようになる。文字列の先頭から順に1文字ずつこのカテゴリ単語辞書113と照合し、終端記号◎までたどり着くような文字列を単語として許容し、出力する。図23では上記の7つの単語のみを照合するようなトライを示しているが、実際は全単語について代表文字コード列に変換してトライを生成し、これをカテゴリ単語辞書113とする。品詞や文字単語などの対応する情報は、終端記号の部分に対応づけておけばよい。あるいは、図22に示すような表とともに、図23に示すようなトライによる辞書を併せ持っていてもよい。もちろん、他のデータ構造によってカテゴリ単語辞書113を構成してもよい。

【0092】また図24は、本発明の文書処理装置の第2の実施の形態におけるコード変換テーブルの一例の説明図である。このコード変換テーブル114は、特に代表文字コード列と、その代表文字コード列に対応する単語を組にして記憶している。ここでは品詞の情報も付加されている。カテゴリ単語辞書113とコード変換テーブル114の保持するデータはほぼ同じであるので、実際の処理では共有することが可能である。しかし、ここでは説明を簡単にするために、別々のデータとして扱うことにする。

【0093】以上の処理は、文書画像から単語を切り出すために必要なデータを準備するための処理であるので、別の装置上で以上の処理を行ない、得られた類似文字カテゴリテーブル、文字コード・カテゴリ対応テーブル、カテゴリ単語辞書113、コード変換テーブル114を予め作成し、それぞれのデータのみを使用するようにしてもよい。

【0094】次に、文書の登録処理について説明する。テキスト領域抽出部104は、画像入力部101で入力された2値のディジタル画像を解析して文字領域を切り出し、さらに各文字を切り出す。このテキスト領域抽出部104の処理は、上述の第1の実施の形態における擬似文字認識部12の処理の一部、すなわち図8に示したフローチャートにおけるS51，S52の処理と同じであるので、ここでは説明を省略する。

【0095】擬似文字認識部105は、テキスト領域抽出部104で切り出された文字領域ごとに処理を行なう。この擬似文字認識部105の処理は、上述の第1の実施の形態における擬似文字認識部12の処理の一部、すなわち図8に示したフローチャートにおけるS53以

降の処理を行なうが、Ｓ５４におけるｂｉ－ｇｒａｍテーブルへの登録処理は行なわない。

【００９６】擬似文字認識部１０５は、テキスト領域抽出部１０４で切り出された各文字画像ごとに、類似文字カテゴリの代表文字コードへ変換する。この処理は、上述の図１０に示した処理と同じであるので説明を省略する。得られた代表文字コード列は、文字ブロックごとにその画像上の位置と入力画像とともに擬似文字認識結果記憶部１１２に記憶される。例えば上述の図１３に示したように、代表文字コードと画像上で占める矩形位置を（左上Ｘ座標，左上Ｙ座標，幅，高さ）で表現し、記憶しておくことができる。

【００９７】カテゴリ単語検出部１０６は、擬似文字認識部１０５で擬似文字認識結果記憶部１１２に格納された代表文字コード列から、カテゴリ単語辞書１１３との照合を行なって、単語として認定される代表文字コード列を抽出する。図２５は、図２６は、本発明の文書処理装置の第２の実施の形態におけるカテゴリ単語検出部の動作の一例を示すフローチャートである。なお、ここではカテゴリ単語辞書１１３は、図２３で示したトライのデータ構造を有しているものとする。

【００９８】まずＳ１２１において、擬似文字認識部１０５で検出した句読点をカテゴリ文字列から検出し、先頭文字から句読点、あるいは句読点間の代表文字コード列を１つの処理ユニットとして、擬似文字認識結果記憶部１１２に記憶されている代表文字コード列を複数の処理ユニットに分割する。以下、分割した処理ユニットを順に処理してゆく。

【００９９】Ｓ１２２において、未処理の処理ユニットがあるか否かを判定し、すべての処理ユニットが処理済みであれば、カテゴリ単語検出部１０６の処理を終了する。未処理の処理ユニットが存在する場合には、Ｓ１２３において、ある未処理の処理ユニットを特定し、その処理ユニットの文字数を変数Ｎに格納するとともに、変数Ｉ，Ｊの値を１にセットする。変数Ｉは処理ユニット中の処理対象の文字を示すために用いられる。また、変数Ｊは、カテゴリ単語辞書１１３内のトライのノードの階層を示すために用いられる。また、Ｓ１２４において、変数Ｐに変数Ｉの値を代入するとともに、変数Ｔの値を１にセットし、領域ＢＵＦＦＥＲをＮＵＬＬにクリアする。変数Ｐは、選択した処理ユニット内で新たな単語の検出を開始した文字位置を示し、変数Ｔは新たに検出した単語の文字数を示すために用いられる。領域ＢＵＦＦＥＲには、検出した単語が順に格納される。

【０１００】Ｓ１２５において、選択した処理ユニットのすべての文字が処理されたか否かを、変数Ｉが変数Ｎ以内か否かで判定する。未処理の文字が存在する場合には、Ｓ１２７において、選択された処理ユニットのＩ番目の文字を、カテゴリ単語辞書１１３のＪ番目の階層の全てのノードの内、処理ユニットのＩ－１番目の文字と接続性のある全てのノードと照合する。このとき、終端記号は全ての文字と一致するワイルドカードとして扱う。ここで、Ｓ１２８において照合する文字があるか否かを判定し、照合する文字がなければ、Ｉ番目までの文字列は単語として許容されないので、現在の単語照合を開始した文字の次の文字から照合をやり直す。すなわち、Ｓ１３２において変数Ｐに格納されている単語の検出開始文字位置に、変数Ｔに格納されている新たに単語として検出された文字数を加算して、これから照合を開始する文字位置を計算して変数Ｉに代入する。さらにＳ１３４において、領域ＢＵＦＦＥＲに格納されているそれまでに検出した単語を記憶装置１０９に格納し、Ｓ１３５でカテゴリ単語辞書１１３の最初から照合を行なうように変数Ｊを１にセットして、Ｓ１２４へ戻る。Ｓ１２４で変数Ｐに変数Ｉの値が代入され、新たに単語の検出を開始する位置を待避する。そして、新たに単語を検出すべく、処理を続ける。

【０１０１】Ｓ１２８で照合した文字が、カテゴリ単語辞書１１３のＪ番目の階層の処理ユニットのＩ－１番目の文字と接続性のあるノードとして存在する場合、さらにＳ１２９において照合した文字に終端記号が含まれるか否かを判定する。終端記号が含まれる場合、変数Ｉが示す文字位置までに単語が含まれる可能性があるので、Ｓ１３０において、検出された単語を領域ＢＵＦＦＥＲに記憶し、単語長を変数Ｔに記憶する。

【０１０２】Ｓ１２７で照合した際に一致する文字は１つに限らず、例えば、ある文字と終端記号の２つと一致する場合がある。Ｓ１３１において、照合により一致した文字が終端記号だけであったか否かを判定し、終端記号だけであれば、それ以上の長さの一致する単語はカテゴリ単語辞書１１３中に存在しないので、Ｓ１３４において領域ＢＵＦＦＥＲに記憶されている、それまでに検出した単語を記憶装置１０９に格納し、新たな単語を検出すべく、Ｓ１３５で変数Ｊの値を１にセットしてカテゴリ単語辞書１１３の先頭に階層を戻し、Ｓ１２４に戻る。Ｓ１２４で変数Ｐに変数Ｉの値が代入され、新たに単語の検出を開始する位置を待避する。そして、新たに単語を検出すべく、処理を続ける。

【０１０３】Ｓ１２９において照合により一致した文字中に終端記号を含まない場合、あるいは、Ｓ１３１において照合により一致した文字が終端記号のみでなかった場合には、Ｓ１３３で次の文字の照合を行なうべく、変数Ｉ，Ｊの値に１だけ加算し、Ｓ１２５へ戻る。

【０１０４】このような処理を繰り返してゆくと、終端記号が現われるごとに単語が検出されて記憶装置１０９に格納されてゆく。そして選択した処理ユニットのすべての文字について処理が終わると、それをＳ１２５において検出し、領域ＢＵＦＦＥＲに格納されている単語を記憶装置１０９に格納して、その処理ユニットについての処理を終了する。Ｓ１２２で未処理の処理ユニットが

存在すると判定された場合には、その未処理の処理ユニットを選択し、上述のように1文字ずつ照合処理を行ない、単語を検出してゆく。すべての処理ユニットについて処理が終了すると、カテゴリ単語検出部106の処理を終了する。

【0105】具体例として、例えば図23に示したトライを用いて代表文字コード列「父化送琴」を処理ユニットとした照合を考える。最初に「父」の照合を行ない、一致するので次の代表文字コード「化」の照合を行なう。図23に示したトライの2番目の階層の「父」と接続性のあるすべての代表文字コード「手」、「羊」、「化」と照合を行なう。この照合により「化」が一致する。一致する代表文字コードの中に終端記号が含まれないので、さらに次の代表文字コード「送」についての照合を行なう。すなわち3番目の階層の代表文字コードの中で「化」と接続性のある終端記号、「送」、「郵」との照合を行なう。この場合、終端記号と「送」と一致する。終端記号を含むので「父化」が単語として検出され、領域BUFFERに格納される。しかし、一致した代表文字コードは終端記号だけではなかったので、さらに照合と続ける。次の代表文字コード「琴」と、4番目の階層の代表文字コード中の「送」と接続性のある「屋」との照合を行なう。しかし代表文字コードは一致しないので、領域BUFFER内の単語「父化」が記憶装置109に格納される。

【0106】次の単語の照合は、検出したカテゴリ単語「父化」の次の文字「送」から始められる。このような処理を処理ユニットの最後の文字まで行ない、さらに未処理の処理ユニットがなくなるまで続ける。この処理により、カテゴリ単語辞書113内に存在し、文書中に出現する全てのカテゴリ単語を記憶装置109に格納することができる。

【0107】文書中には同一の単語が複数回出現するのが一般的であるため、記憶装置109には同じカテゴリ単語が重複して格納されることになる。重複したカテゴリ単語は、そのまま残してもよいし、1つ以外を削除してもよい。画像上での単語の出現場所を知りたい場合のために、擬似文字認識結果記憶部112に代表文字コードとともに画像上の位置情報を記憶しているが、重複排除する場合には、1つの単語に複数の位置情報を記憶するように構成すればよい。なお、この位置情報を用いて単語の出現位置を表示する等の手法としては、周知の種々の技術を用いることができ、ここではこれ以上の説明は行なわない。

【0108】これまでの処理で、代表文字コードで表現されたカテゴリ単語を抽出することができた。しかし、これまでの処理では単に単語辞書に存在する単語を検出しているのみで、必ずしも日本語の単語として許容できるものである保証はない。例えば、複合名詞を本来の名詞の境界とは違う文字で分けて単語を抽出したり、付属語との接続が誤っている単語を抽出する可能性がある。そのため、以下で説明するように、単語間の品詞の接続性を検証することで、このような言語として誤りを訂正する。

【0109】例えば「将来、実現される技術である。」という文を代表文字コード列で表わすと、例えば「均糸。芸温される転術である。」となる。このうちの処理ユニット「芸温される転術である」について、上述のようにカテゴリ単語の検出を行なってゆくと、例えば、図22に示したカテゴリ単語辞書113からカテゴリ単語「芸温」が検出され、さらに「さ」、「れる」が検出される。カテゴリ単語「芸温」は、文字単語「実現」と「差損」を、また、「さ」は「さ」（サ変動詞の語尾）と「き」（下一段活用動詞の語幹）をそれぞれ含んでいる。しかし、文脈を考えると、「実現」あるいは「差損」という名詞の後に「き」という語幹を有する動詞が続くのは、文法上おかしい。また、「差損」という名詞には、使役の助動詞が続くことはない。したがって、「実現」と「さ」の単語の組合せの解釈が正しい。同様に、「さ」と「れる」の組み合わせも正しい。なお、この場合の実際の品詞は、サ変動詞「実現する」と使役の助動詞「れる」である。

【0110】このような単語抽出の誤りは、普通の文字列の解析においても発生するが、曖昧性の多い代表文字コード列の方が発生しやすいと考えられる。そのため、単語を検出するたびに先に検出した単語との接続性を検証することで、抽出する単語の精度を向上させることができる。

【0111】このような接続性の検証に、カテゴリ単語辞書113に格納されている品詞接続辞書を用いることができる。図27は、本発明の文書処理装置の第2の実施の形態における品詞接続辞書の一例の説明図である。図27に示した品詞接続辞書は、連続する2つの単語の品詞の接続関係を示したもので、先の単語の品詞を行に、後の単語の品詞を列に対応づけて示している。表の値は、

$L_{ij}=1$　；　行iの品詞は列jの品詞に接続可能

$L_{ij}=0$　；　行iの品詞は列jの品詞に接続不可能

であることを示している。カテゴリ単語を検出する度に、例えば図27に示すような品詞接続辞書を用いて単語間の接続関係を検証する。

【0112】しかしながら、カテゴリ単語は1つの代表文字コード列で複数の文字単語を表現する可能性がある。したがって、実際の処理では、単語として抽出された代表文字コード列に対応する複数の文字単語の品詞全てに対して接続性の検証を行ない、そのうちの1つでも接続性が検証されれば、その代表文字コード列を単語として認定する。

【0113】図28は、本発明の文書処理装置の第2の実施の形態における品詞接続関係の検証処理の一例を示

すフローチャートである。この処理の入力はカテゴリ単語検出部106で検出されるカテゴリ単語であり、検出されるたびに順次入力されて単語間の接続性を検証するものである。まず、S141において、処理ユニットで最初に検出されたカテゴリ単語を入力し、変数WORD1に代入する。そしてS142において、このカテゴリ単語が取り得る品詞が、文節の先頭となりえるか否かを調べる。このとき、カテゴリ単語が文節の先頭となりえる品詞の単語を含んでいなければ、このカテゴリ単語は日本語として受理できないので、この代表文字コード列を単語として拒絶する。

【0114】S142でカテゴリ単語が文節の先頭となりえると判定された場合、S143において、処理ユニットから次のカテゴリ単語を入力し、変数WORD2に格納する。そしてS144において、変数WORD1と変数WORD2に格納された2つのカテゴリ単語の接続性を、例えば図27に示したような品詞接続辞書を検索して求める。このとき、2つのカテゴリ単語の取りえる品詞のうち、接続関係の成立する品詞の組合せがない場合、変数WORD1に格納されている最初のカテゴリ単語は日本語として受理できないため、最初の単語は拒絶される。2つのカテゴリ単語の取りえる品詞の組合せのうち、接続関係の成立する品詞の組合せが存在している時、S145において変数WORD1に格納されている最初のカテゴリ単語を正しい単語として受理する。さらに、S146において変数WORD2に格納されているカテゴリ単語を変数WORD1に移す。S147で処理ユニットの終端か否かを判定し、終端でなければS143に戻り、処理ユニットの残りのカテゴリ単語を順に入力して、同様に単語間の品詞の接続性を検証する。

【0115】S145で受理されたカテゴリ単語は、どの品詞でカテゴリ単語が受理されたかを記憶装置109に記憶しておくことで、あとのカテゴリ単語変換部107での出力をより正確にすることができる。単語として拒絶された場合は、現在処理中の文節の先頭文字まで戻り、再度、カテゴリ単語検出部106で単語の抽出を行ない、別の単語候補を抽出する。

【0116】具体例を用いて、上述の処理の流れを説明する。ここでは先に示した例「将来、実現される技術である。」を用いて、「実現される」という文節内の品詞を決定する様子を説明する。まず、代表文字コード列「芸温される」の先頭の文字から順にカテゴリ単語辞書113と照合し、単語の可能性のある代表文字コード列「芸温」を得る。図22に示したカテゴリ単語辞書113の内容から、この代表文字コード列は「差損」（名詞）、「実現」（名詞）、「実現」（サ変動詞語幹）のいずれかである可能性がある。次に検出される代表文字コード列は「さ」であり、「さ」（サ変動詞語尾・未然形）または「き」（下一段動詞語幹）である可能性がある。図27に示した品詞接続辞書を参照すると、（名詞）－（サ変動詞語尾）、（名詞）－（下一段動詞語幹）の接続関係は存在しないので、この時点で最初の単語の可能性のある代表文字コード列「芸温」は「実現」（サ変動詞語幹）であることが分かる。そのため、代表文字コード列「芸温」がカテゴリ単語として受理される。

【0117】次に検出される代表文字コード列は「れる」（助動詞）であるが、図27に示した品詞接続辞書によると、語尾・未然形と接続可能なことが分かる。したがって、代表文字コード列「さ」はサ変動詞の活用語尾であることが決定され、カテゴリ単語として受理される。さらに、図27に示した品詞接続辞書によると、助動詞は文節終端となりえるので、代表文字コード列「れる」もカテゴリ単語として受理され、「芸温される」という代表文字コード列は1つの文節として受理されることになり、サ変動詞の語幹「実現」が自立語として検出される。

【0118】また、単語の切り出し位置が間違っており、品詞の接続性が検証されない場合には、処理を行なっている文節の先頭に戻り、切り出し位置を変えて、再度、品詞の接続性の検証を行なう。以上の方法により、複合名詞の単語の境界、あるいは言語的な単語の接続性を保ち、単語を抽出することができる。

【0119】以上の処理により、少なくとも日本語として受理できる接続関係をもつカテゴリ単語が検出された。次に受理されたカテゴリ単語を通常の文字で構成される単語に変換する。この処理は、カテゴリ単語変換部107でコード変換テーブル114を用いて行なわれる。カテゴリ単語変換部107での処理は簡単である。単語として受理されたカテゴリ単語をコード変換テーブル114で検索し、対応するカテゴリ単語が取りえる全ての文字単語を出力する。ただし、検索に用いられる単語は自立語なので、先の品詞接続関係の検証処理で、自立語として認定された単語だけを出力する。これにより、検索に有効な単語を得られるとともに、出力される単語数を抑制することができる。

【0120】以上のようにして、文書画像から文字認識処理のような計算機パワーを必要とする処理を用いることなく、単語を抽出することができる。ここで検出された単語は、品詞の接続関係を検証しているために、文章として成立しないような単語は含まれていないので、検索に用いた場合、精度の高い検索が期待できる。なお、この第2の実施の形態では、カテゴリ単語変換部107で代表文字コード列を文字コード列に変換しているので、上述の第1の実施の形態に示したように検索式中のキーワードを代表文字コード列に変換することなく、そのまま文字コード列によって検索を行なうことができる。

【0121】次に、本発明の文書処理装置の第2の実施の形態における第1の変形例について説明する。ここで

は、上述の第1の実施の形態における第1の変形例と同様に、擬似文字認識部105で文字画像を代表文字コードに変換する際に、その精度を向上させた例を示している。上述の例では、文字画像に対して各類似文字の代表文字コードを割り当てる際に、図10のS65で説明したように、特徴空間で特徴量が最も近いものを選択する最短距離識別法を用いている。しかし、実際の文字画像の特徴量は、画像のかすれや歪みにより、類似文字のクラスタは互いに複雑に重複していることが多い。この場合、最短距離識別法では誤識別を起こす可能性が高い。

【0122】図29は、代表文字コードの誤識別の一例の説明図である。例えば、ある2次元の特徴量による空間において、図29に示すように2つのクラスタa1とa2が存在する場合を考える。xという未知文字は、本来クラスタa1に属する文字である。しかし、最短距離識別法では、未知文字xは距離の最も近いクラスタa2に属していると判定される。このような誤識別は、2つのクラスタが重複しているとき、未知文字xの特徴量が2つのクラスタの共通部分に存在する場合にも同様に発生する。

【0123】このような誤識別の問題を解決するために、上述の第1の実施の形態における第1の変形例では、ε-component拡張法を用いて、1文字種を複数の類似文字カテゴリに登録している。このようにして生成された類似文字カテゴリテーブルを用いてカテゴリ単語辞書113を生成すると、1文字種が複数のカテゴリに属しているため、複数の異なる代表文字コード列が同じ1つの文字単語を表わすことになる。例えば、文字「画」がカテゴリ「画」に、文字「像」がカテゴリ「倶」と「根」に登録されていると、単語「画像」は、カテゴリ単語「画倶」と「画根」という2つの代表文字コード列で表されることになる。このように1つの文字単語に複数の異なる代表文字コード列が対応すると、結果としてカテゴリ単語辞書113のサイズを増大させることになる。このような、辞書サイズの増大は、カテゴリ単語辞書113の構成を複雑にするだけでなく、単語の抽出速度にも影響を与える。

【0124】そのため、ここでは、類似文字分類部103では最短距離識別法を用いて、類似文字カテゴリテーブル41を生成し、これまでと同じカテゴリ単語辞書113を生成し、擬似文字認識部105での識別時には、入力された文字画像の特徴量と各類似文字カテゴリのカテゴリ代表ベクトルとのユークリッド距離を計算して、その距離の近い方からN番目までのカテゴリを入力文字の文字カテゴリとして代表文字コードを出力する。ただし、距離に閾値Dtを設けて、閾値Dt以上離れている文字カテゴリは、入力文字の文字カテゴリに採用しないようにして、1文字種に対応する文字カテゴリを絞りこむこともできる。

【0125】図30は、本発明の文書処理装置の第2の実施の形態の第1の変形例におけるN=2とした場合の代表文字コード列への変換の一例の説明図である。例えば、上述の方法で「自然言語処理」という文字列を代表文字コード列に変換する場合を考える。ここでN=2とした。また、「語」の文字には閾値Dt以内に最短距離に存在する類似文字カテゴリのみが存在しているとする。

【0126】N=1、すなわち最短距離識別法により変換された代表文字コード列は「自滅豆記肋喫」である。例えば、第3文字目の代表文字コード「豆」のカテゴリには、文字「言」が含まれていないものとする。このとき、「自滅豆記肋喫」という代表文字コード列からは、文字列「自然言語処理」を再現することはできない。

【0127】N=2までの代表文字コード列を考える。すなわち、距離が閾値Dt以内で次に距離的に近いカテゴリを求める。これにより、文字「自」については代表文字コード「吉」が、「然」については「恩」が、「言」については「吉」が、「処」については「近」が、「理」については「均」が、それぞれ得られる。文字「言」が、このようにして得られた代表文字コード「吉」のカテゴリに含まれていれば、文字列「自然言語処理」を再現することが可能となる。

【0128】このようにして1つの文字に対して1以上得られた代表文字コードからなるカテゴリ文字列から、カテゴリ単語検出部106で単語を抽出する。カテゴリ単語検出部106での処理は、上述の方法を変更することなく、全ての代表文字コードをカテゴリ単語辞書113と照合して、単語として許容できる代表文字コード列を記憶装置109に記憶する。すなわち、第1文字目に「自」または「吉」という代表文字コードを取り、それに続く第2文字目に「滅」または「恩」という代表文字コードが続くか否かを照合する。以下同様に終端記号を検知するまで照合を続け、終端記号を検出したところで、それまで検出した代表文字コード列をカテゴリ単語として記憶装置109に記憶する。このとき、処理途中で複数のカテゴリ文字列が生成されるが、単語として続く文字が存在しないところでその代表文字コード列は棄却すればよい。

【0129】例えば、第2文字目までの照合で、「自恩」、「自滅」、「吉滅」の3つの単語候補が存在しているとする。この時点で、カテゴリ単語辞書113上で終端記号を検出して、単語として認定されている代表文字コード列は「自滅」、「自恩」の2つであるとする。次に続く代表文字コードは「豆」または「吉」であるが、「自恩」-「豆」あるいは「自恩」-「吉」と続く単語が、照合用のカテゴリ単語辞書に存在しなければ、以後の照合では、「自恩」で始まるカテゴリ単語の照合は行なわない。次に4文字目「記」を照合すると、「吉滅豆」-「記」あるいは「吉滅吉」-「記」と続く単語が、照合用の単語辞書に存在しない場合、これまでの照

合では、カテゴリ「吉」で始まる代表文字コード列は終端記号と照合して、単語として認定されている単語が存在しないので、以後の照合処理では、カテゴリ「吉」で始まる単語の候補を棄却する。さらに、処理を続け、第7文字目まで照合すると、代表文字コード列「自滅吉記肋喫」に続く文字は終端記号のみであった場合、代表文字コード列「自滅吉記肋喫」を単語として認定する。

【0130】ここで、第1文字目「自」あるいは「吉」から始まり、照合用の単語辞書で単語として認定された単語は、「自恩」と「自滅吉記肋喫」である。しかし、ここでは最長一致の原則を用いて、長い単語として検出された代表文字コード列「自滅吉記肋喫」のみを単語候補として残し、出力する。また、上述したように、検出された単語が言語として許容できるかを、品詞接続辞書との照合で検証し、言語として許容できるカテゴリ単語列のみを出力する。

【0131】以上のように、1つの文字画像に対して、複数の類似文字カテゴリを対応つけることで、より正確に単語の抽出を行なうことが可能となる。このように、文字画像のかすれや歪みによる文字特徴の変動により、最短距離識別によるカテゴリ文字の選択では擬似文字認識で誤りを起こす文字を、特徴の近い文字カテゴリを複数選択することにより、擬似文字認識の誤りを最小限に止めることができる。

【0132】次に、本発明の文書処理装置の第2の実施の形態における第2の変形例について説明する。上述のように、この第2の実施の形態およびその第1の変形例では、全文字種に対する詳細な識別処理を行なうことなく、文書画像中から言語として許容できる単語を抽出することができる。しかしながら、これまでは類似文字カテゴリの組合せとして単語を抽出しているので、曖昧性が残り、1つの単語として抽出された代表文字コード列に複数の文字単語が対応する場合がある。例えば、名詞として許容された「琴記」というカテゴリ単語は、「単語」と「筆記」の2つの文字単語が対応する。上述の構成では、「単語」と「筆記」の2つの単語を文書画像に書かれている自立語として抽出することになり、いずれの単語が文書画像中に記述されているかを判別することはできない。

【0133】このような問題を解決するために、この第2の変形例では、各文字の特徴を詳細に調べて、文字を一意に決定する。この場合、従来の文字認識のように約3000文字種に対して、特徴量の比較を行なう必要はなく、カテゴリ単語検出部106で検出されたカテゴリ単語に対応する文字単語で使用されている文字との特徴量の比較で済む。例えば、検出されたカテゴリ単語を3つの文字単語と解釈できる時、詳細識別処理では、各文字位置で3文字との特徴量の比較を行なえばよいことになる。

【0134】図31は、本発明の文書処理装置の第2の実施の形態における第2の変形例を示す構成図である。図中、図21と同様の部分には同じ符号を付して説明を省略する。110は詳細識別部、115は詳細識別辞書である。詳細識別部110は、入力された未知文字の詳細な特徴量を抽出して、類似文字カテゴリ内の文字の特徴量と比較し、文字種を一意に決定する。詳細識別辞書115は、類似文字カテゴリごとに文字画像の詳細な特徴を記憶する。

【0135】詳細識別部110と詳細識別辞書115についてさらに説明する。詳細識別辞書115は、類似文字分類部103で類似文字に分類された結果である類似文字カテゴリテーブルを用いて作成される。詳細識別辞書115を作成するために用いられる特徴量は、従来の文字認識装置で用いられている特徴量を使用することができる。図32は、本発明の文書処理装置の第2の実施の形態の第2の変形例において詳細識別辞書を作成するために用いる特徴量の一例の説明図である。使用する特徴量として、例えば、特開平5－166008号公報に記載されている方向属性を用いた特徴量を適用することができる。この特徴量は、文字画像中の輪郭画素に対して、画素の連続性を複数の方向について計測したもので、文字を構成する線分の方向や複雑さを表わしている。図32（A）に示した例では、「漢」という文字画像の輪郭画素について、それぞれ左右方向、上下方向、左上－右下の斜め方向、右上－左下の斜め方向について連続性を示す画素数を計数し、最も計数値の大きい方向を求めてその画素の方向属性とする。左右方向に最も計数値が大きくなる輪郭画素を集めると図32（B）に示す特徴が得られる。同様に、上下方向に最も計数値が大きくなる輪郭画素を集めると図32（C）に示す特徴が得られ、左上－右下の斜め方向では図32（D）、右上－左下の斜め方向では図32（E）に示す特徴が得られる。このような方向属性の特徴を詳細識別辞書として格納しておけばよい。

【0136】また、萩田他，「外郭方向寄与度特徴による手書き漢字の識別」，電子情報通信学会論文誌D，Vol．J66－D，No．10，pp1185－1192，1983年10月で提案されている外郭方向寄与度特徴を用いてもよい。類似文字分類部103で用いているペリフェラル特徴が文字の外形を表わすのに対して、これらの特徴量は、いずれも文字内部の線の複雑さ、方向、接続性を表わし、文字のより詳細な特徴を表現している。もちろん、その他の特徴を用いても、複数の特徴量を組み合せて用いてもよい。

【0137】図33は、本発明の文書処理装置の第2の実施の形態の第2の変形例における詳細識別辞書の作成手順の一例を示すフローチャートである。なお、ここでは使用する特徴量を詳細特徴として表現し、特定の特徴量として述べることはしない。まずS151において、類似文字カテゴリテーブルから1つの類似文字カテゴリ

を選択する。次にＳ１５２において、トレーニングサンプルの画像から、選択した類似文字カテゴリに属している文字種を表わす画像を取り出す。Ｓ１５３において、Ｓ１５２で取り出した文字画像から文字種ごとに詳細特徴を抽出し、Ｓ１５４において、詳細特徴の平均を算出する。Ｓ１５５において、この特徴量を類似文字カテゴリごとにまとめて詳細識別辞書１１５に追加してゆく。このような処理を各類似文字カテゴリごとに行なうことによって、詳細識別辞書１１５を生成する。

【０１３８】図３４は、本発明の文書処理装置の第２の実施の形態の第２の変形例における詳細識別辞書の一例の説明図である。詳細識別辞書１１５は、例えば図３４に示すように、類似文字カテゴリごとに、そのカテゴリに属する文字コードとその詳細特徴量ベクトルにより構成することができる。この詳細識別辞書１１５は、類似文字カテゴリテーブルや、文字コード・カテゴリ対応テーブル、カテゴリ単語辞書１１３、コード変換テーブル１１４と同様に、別の装置上で予め用意しておいて、それぞれのデータのみを使用するように構成することも可能である。

【０１３９】上述のように、この第２の実施の形態では、言語として許容できるカテゴリ単語を代表文字コード列から抽出し、カテゴリ単語変換部１０７により、最終的に文字単語を得ている。このとき、１つのカテゴリ単語に対して、複数の文字単語への変換が可能である場合がある。このような時、詳細識別部１１０を呼び出して、各文字画像を詳細に識別し、一意に文字コードを決定して文字単語を決定する。

【０１４０】詳細識別部１１０では、以下のような処理により文字単語を決定する。いま、複数の単語に変換可能なカテゴリ単語をＳｃとし、カテゴリ単語Ｓｃの長さをＬ（Ｓｃ）で表わす。また、カテゴリ単語Ｓｃが変換可能な文字単語数をＮとし、第ｎ（≦Ｎ）番目の候補単語をＳｗｎとする。ただし、候補単語として順番をつけているが、番号が若いほど単語として成立しやすいなどの意味はなく、単に辞書順で便宜上番号付けを行なっている。さらに、文字単語Ｓｗｎのｉ番目の文字を同様にＳｗｎ（ｉ）と表わす。ここで、入力された未知文字Ｘと、ある文字Ｍとの特徴量の差をＦ（Ｘ，Ｍ）とするとき、

$$An=\sum_{i=0}^{L(Sc)} F(X(i), Swn(i))$$

なる式の値Ａｎが最小となる文字単語をカテゴリ変換部１０７の最終的な結果として出力する。

【０１４１】この式から分かるように、実際には詳細識別部１１０では、各カテゴリ内の全ての文字種との比較を行なう必要はなく、単語として可能性のある候補単語内の文字種とのみ比較を行なえばよい。最悪の場合でも、１カテゴリに対する比較回数は、カテゴリ内の類似文字数である。

【０１４２】ここで、特徴量の差の累積値を用いているのは、各文字画像の詳細識別を行なって、各文字ごとに最も確からしい文字を組み合せて単語を作った場合に、文字画像のかすれや歪み等の影響で、候補単語以外の単語（ときには、言語として許容できない単語）を生成することが考えられるからである。少なくとも、カテゴリ単語検出部１０７で検出されている単語は、言語的には許容されている単語なので、カテゴリ単語検出部１０７で検出された候補単語だけを識別対象とすることができる。

【０１４３】図３５は、本発明の文書処理装置の第２の実施の形態の第２の変形例における詳細識別部の処理の一例を示すフローチャートである。上述の詳細識別部１１０における処理の一例を、図３５を用いてさらに説明する。まず、Ｓ１６１において、処理対象となるカテゴリ単語Ｓｃを選択し、そのカテゴリ単語Ｓｃに対応する文字単語の候補数Ｎを計数する。また、そのカテゴリ単語Ｓｃの長さＬ（Ｓｃ）をＷとする。さらに、処理に使用する記憶領域Ａ［Ｎ］の確保と初期化を行なうとともに、変数ｉを１に初期設定する。このとき、文字単語の候補数Ｎが１のときは、処理対象から外され、そのままカテゴリ単語変換部１０７により文字単語へ変換が行なわれる。そして、処理対象となったカテゴリ単語Ｓｃを文書画像中から切り出す。処理対象のカテゴリ単語Ｓｃの文書画像中での位置は、カテゴリ単語検出部１０６でカテゴリ単語を切り出す際に位置情報を保存しておき、これを参照することで知ることができる。

【０１４４】次に、Ｓ１６２において第ｉ文字目の文字画像を切り出す。カテゴリ単語内の各文字画像の位置は、擬似文字認識部１０５において各文字画像を文字カテゴリに割当てる際に、同時に位置情報を保存しておき、これを参照することで知ることができる。このようにして切り出した文字画像から、Ｓ１６３において、詳細識別辞書１１５を作成した時と同じ特徴量を抽出する。これを特徴量Ｘとする。Ｓ１６４～Ｓ１６７において、抽出した特徴量と各候補単語の第ｉ文字目の詳細特徴量とを比較し、その差を候補単語ごとに記憶領域に累積する。すなわち、Ｓ１６４で変数ｊを１にセットし、Ｓ１６５において、Ｓ１６３で抽出した特徴量Ｘと第ｉ文字目の詳細特徴量Ｓｗｊ（ｉ）の特徴量の差Ｆ（Ｘ，Ｓｗｊ（ｉ））を計算し、Ａ［ｊ］に累積する。Ｓ１６６で変数ｊを１だけ増加させ、Ｓ１６７で変数ｊの値が文字単語候補数Ｎを越えたか否かを判定し、越えるまでＳ１６５に戻って処理を続ける。これにより、記憶領域Ａ［１］～Ａ［Ｎ］にそれぞれ第１～ｉ文字目までの特徴量の差が累積される。

【０１４５】さらに、Ｓ１６８で変数ｉに１を加え、Ｓ１６９でカテゴリ単語の長さＷと比較して変数ｉの値がＷ以下の場合にＳ１６２へ戻り、処理を続ける。このようにして、最後の文字までＳ１６２～Ｓ１６９の処理を繰り返すことによって、記憶領域Ａ［１］～Ａ［Ｎ］に

は、各文字単語候補ごとに、各文字の特徴量の差の累積値が格納される。

【0146】S170において、記憶領域A［1］～A［N］の値を比較し、最小値を持つ記憶領域のアドレスCを求める。S171において、このアドレスCに対応する候補単語SwCを抽出し、その単語を最も確からしい文字単語として出力する。

【0147】ここでは未知文字の特徴量と辞書の特徴量との差の累積値を単語の評価関数として用いた例を示したが、辞書作成時に得られるトレーニングサンプルの詳細特徴量の分散等の統計的な情報を用いて、統計的に未知文字の確からしさを求めて、その値の累積を単語の評価関数としてもよい。

【0148】以上のように、カテゴリ単語検出部107で検出したカテゴリ単語を複数の文字単語に変換可能な時、検出したカテゴリ単語に対して詳細識別を行なうことで、正確に単語を抽出することができる。また、詳細識別の対象を候補単語の文字の組合せに限定することで、カテゴリ単語検出部107で検出した、言語的に許容できる単語を得られることが保証される。

【0149】次に、本発明の文書処理装置の第2の実施の形態における第3の変形例について説明する。第2の実施の形態における上述の各例では、文字切り出しの段階での誤りがないものとしてきた。しかし、上述の第1の実施の形態の第2の変形例でも説明したように、切り出し段階での誤りは、現実には多く存在する。この第3の変形例では、このような切り出しの誤りに対応する例を示す。ここでは一例として、上述の第1の実施の形態の第2の変形例と同様、図16に示した例について考える。

【0150】図36は、本発明の文書処理装置の第2の実施の形態の第3の変形例における切り出された文字列の関係の一例の説明図である。上述のように、図16（A）に示した「文書印刷」の例の場合、「文」、「書」については文字間の間隙しか存在しないが、「印」の文字中に1か所、「刷」の中に2か所、垂直方向に白画素のみからなる切り出し位置候補が存在するとともに、これら2文字の間も当然切り出し位置が存在するので、合計5つの部分文字（a1，a2，b1，b2，b3）が得られる。これらについて、文字としての統合を試みる。文字「文」、「書」と統合できるものはないので、そのまま1文字として、擬似文字認識部105において類似文字カテゴリの識別を行ない、代表文字コード「父」、「君」に変換される。文字「印」については、部分文字a1，a2を2つの文字として扱う場合と、1つの文字として扱う場合の2つの可能な解釈がある。a2とb1を統合した場合は幅のしきい値を越えるため、統合はなされない。したがって、ここまでの2つの解釈を同じ文字画像領域に対して保持する必要がある。これらそれぞれについて、擬似文字認識部105において類似文字の識別を行なうと、部分文字a1は「E」、部分文字a2は「P」、a1a2では「叩」という代表文字コードに変換され、記憶装置109に格納される。図36において、代表文字コードを括弧書きで示している。また、図中の○は文字切り出しの解釈の区切りである。

【0151】同様にb1以降を順に見ていくと、可能な解釈が（［b1］，［b2］，［b3］），（［b1b2］，［b3］），（［b1］，［b2b3］），（［b1b2b3］）の4通りあるので（［］は中の部分文字が1つの文字と見なされることを示す）、同様に擬似文字認識部105で処理が行なわれる。［b1］，［b2］，［b3］，［b1b2］，［b2b3］，［b1b2b3］はそれぞれ、「凧」、「1」、「1」、「引」、「リ」、「刷」という代表文字コードに変換される。これらすべての解釈を記憶装置109に保持する。

【0152】このようにして求められた「印刷」に対応する代表文字コード列を、ここでは「［EP，叩］［凧［11，リ］，引1，刷］」のように表現する。［］内は文字画像のある範囲内での切り出し解釈が複数ある場合にそれを並べたものである。これは入れ子にすることができ、例えば「刷」の右部分の2本の垂直ストロークを1つと見なす場合と、2つに見なす場合の2つが表現できる。

【0153】カテゴリ単語辞書113を探索する場合に、複数の切り出し解釈がある場合は、その範囲ごとにそれぞれの代表文字コード列がカテゴリ単語辞書113に存在するか否かを調べ、可能性のあるものはすべて残す。上記の例で、「印」という字に対しては、まず「EP」、「叩」という代表文字コード列がカテゴリ単語辞書113に存在するか否かを調べる。このとき、両者ともに存在するとすれば、両者を存在する可能性のあるものとして保持する。次に文字「刷」に対しては、「EP」、「叩」それぞれについて後に続く代表文字コードとして「凧」、「引」、「刷」があるので、接続可能か否かをカテゴリ単語辞書113で調べる。ここでは、「EP」は3つの候補どれとも接続せず、それ自身で単語となり、「叩刷」の代表文字コード列はカテゴリ単語辞書113中に存在するので、単語として取り出されるので、後に続く単語を同様に照合し、品詞接続辞書による接続性のチェックを行なうことになる。「EP」という解釈については、これをひとつの単語と見なし、次の文字から始まる単語の接続可能性を見る。ここでは接続する可能性のある文字カテゴリは「凧」、「引」、「刷」の3つで、それぞれのカテゴリから始まる単語を取り出し、品詞接続関係を調べる。これらの単語は「EP」との接続するものがないとすれば、「EP」という解釈についての可能性が棄却され、「叩刷」が残ることになる。

【0154】さらに複雑な場合の例として、「NMRにおける」という文字列を考える。ここで、文字「N」、「M」、「R」は半角文字である。そのため、これらの英字については、隣接する英字と統合されて漢字として認識される場合が想定される。さらに、「に」の文字中に切り出し位置が1カ所存在する。

【0155】図37は、本発明の文書処理装置の第2の実施の形態の第3の変形例における切り出された文字列の関係の別の例の説明図である。想定される統合としては、「NM」、「MR」、「R」と「に」の左側のストロークが一つの文字として統合された場合が考えられる。3つの統合文字に対応する代表文字コードとして「肌」、「狼」、「引」が得られたとする。また、文字「に」は「に」という代表文字コードと、分離された部分文字ごとに「1」と「こ」が得られたとする。すると、文字切り出しの複数の解釈を許す代表文字コード列は[N[M[Rに,引こ],[狼[に,1こ]]],[肌[R[に,1こ],引こ]]]のように表現される。実際に照合される場合は代表文字コード列の表記の中で[]で示される複数の切り出し解釈から、代表文字コードをノード、遷移可能な代表文字コード間の接続をアークとするようなカテゴリ遷移データを作成することができる。

【0156】図37には、「NMRにお」という文字列部分を対象にしたカテゴリ遷移データを示している。このカテゴリ遷移データをもとに、先頭からカテゴリ単語辞書113との照合を行なってゆく。例えばカテゴリ単語辞書113中の単語として「NMR」(名詞)、「肌」(名詞)、「肌引」(動詞語幹)という三つが照合されたとする。これから後の単語との接続を品詞接続辞書で照合する。例えば、「NMR」に対しては「に」(格助詞)が接続可能で、「肌」については代表文字コード「R」、「引」からはじまる単語には接続できず、「肌引」に対しても代表文字コード「こ」から始まる単語は接続しないことが分かれば、結果として「肌」、「肌引」という単語候補は棄却され、「NMRに」が候補として残る。このようにして、正しい切り出し位置の候補が残ることになる。

【0157】図38は、図39は、本発明の文書処理装置の第2の実施の形態の第3の変形例における切り出された文字列の統合処理の一例を示すフローチャートである。まず、擬似文字認識部105で処理を行なう単位である処理ユニットの代表文字コード列を、上述のようなカテゴリ遷移データに展開する。S181において、処理ユニットの最初の位置を注目点として処理を開始する。

【0158】S182において、次の代表文字コードへの複数の遷移パスが存在するか否かを判定し、複数の遷移パスが存在する場合には、S183において、参照する階層を1階層深くする。S184において、現在参照している階層において可能な遷移パスを変数Pにセットする。

【0159】S185において、変数Pの中でまだ処理していない遷移パスが存在しているか否かを判定し、未処理の遷移パスが存在する場合、S186において、その中の1つに注目し、遷移パスの先の代表文字コードをカテゴリ単語辞書と照合する。S187で照合に成功したか否かを判定し、照合に成功したらばさらにS188において単語境界か否かを判定する。単語境界でなければS182へ戻り、単語として取り出せるまでカテゴリ単語辞書との照合を行なう。単語境界までの照合が成功し、カテゴリ単語の候補が得られると、S189において、直前に得られたカテゴリ単語候補との品詞の接続関係を調べ、接続可能か否かを判断する。接続可能であれば、それを残してゆく。S190で処理ユニットの最後まで処理したか否かを判定し、処理ユニットの途中であれば次のカテゴリ単語を抽出すべくS182へ戻って処理を続ける。処理ユニットの最後まで処理したらば、それまでに得られたカテゴリ単語の列は1つの候補として成り立つので、S191において得られたカテゴリ単語の列を出力する。

【0160】S186でのカテゴリ単語辞書との照合に失敗したとS187で判定された場合は、S193において、参照する階層を1階層浅くして途中の複数の解釈が存在する位置まで戻り、S185からの他のパスについての照合を行なう。また、S189で品詞の接続が許されていない場合も、それまで仮定してきた単語列の可能性は棄却し、それ以降の遷移は処理対象外として処理は行なわず、S193において参照する階層を1階層浅くして途中の複数の解釈があるところまで戻り、S185からの処理を同様に続ける。さらに、処理ユニットの最後まで処理し、S191でカテゴリ単語の列を出力した後も、他の可能性を判定すべく、S193に進んで1階層浅くして途中の複数の解釈が存在する位置まで戻り、S185に進んで処理を続ける。もちろん、可能性のある他のカテゴリ単語列が得られた場合には、S191で出力される。

【0161】S185で変数Pの中に未処理の遷移パスが存在しなくなると、S192においてトップの階層のすべての遷移を調べたか否かを判定し、調べていない遷移が存在する場合には、S193に進んで1階層浅くして途中の複数の解釈が存在する位置まで戻り、S185で未処理の遷移パスを探して処理を続ける。トップの階層のすべての遷移について処理が終了すれば、1つの処理ユニットについて、与えられたカテゴリ遷移データのすべてのパスについて処理が終了したことを示すので、この統合処理を終了する。

【0162】このように分離した文字があり、複数のカテゴリ単語候補がある場合でも、単語としての可能性を品詞の接続関係から減らしていくことができるので、非

常に高速にかつ精度よく、単語の抽出が可能となる。

【0163】上述の各実施の形態は、コンピュータプログラムによっても実現することが可能である。その場合、そのプログラムおよびそのプログラムが用いる辞書、テーブルなどは、コンピュータが読み取り可能な記憶媒体に記憶することも可能である。記憶媒体とは、コンピュータのハードウェア資源に備えられている読取装置に対して、プログラムの記述内容に応じて、磁気、光、電気等のエネルギーの変化状態を引き起こして、それに対応する信号の形式で、読取装置にプログラムの記述内容を伝達できるものである。例えば、磁気ディスク、光ディスク、CD-ROM、コンピュータに内蔵されるメモリ等である。

【0164】

【発明の効果】以上の説明から明らかなように、本発明によれば、文書画像中の文字を文字コード列にまで識別することなく、少数の類似文字のカテゴリに分類するだけでフルテキスト検索を実現している。本発明での類似文字のカテゴリの識別は、通常の文字認識と比較してはるかに少ない次元の特徴ベクトルを用いて行なっており、少数の類似文字のカテゴリに識別するだけで済むので、文書画像からキーワードとして使用できる自立語の抽出と文書画像登録時の飛躍的な速度の向上が実現できるという効果がある。

【0165】この類似文字のカテゴリを元の文書画像の属性として保持し、検索時には検索キーワードの各文字を類似文字のカテゴリの列に変換して検索する。類似文字のカテゴリには複数の文字が属しているため、キーワードから変換された代表文字コード列に対応する文字列は所望のキーワード以外のものも含まれる可能性がある。しかしながら、検索キーワードは通常複数文字で構成され、しかも複数指定されるということを考えると所望のキーワード以外のものが結果として得られることは現実的には少ない。逆に、類似文字のカテゴリ分けの精度は文書画像中の文字の誤認識等に比べて格段によいので、漏れの少ない検索を実現することができる。また、通常のフルテキストサーチの手法がそのまま使用できるので、通常の電子文書の検索と同様に処理することができるという効果もある。

【0166】また、類似文字のカテゴリの列から、カテゴリ単語辞書をもとに単語として取り出すことによって、無意味な文字列を検索する可能性を減少させ、さらには品詞など単語間の接続性なども考慮することによってさらに検索精度を向上させることができる。同じ類似文字のカテゴリの列によって異なる複数の単語が表現されることもあるが、このような場合にはカテゴリ内のいずれの文字であるかをさらに詳細な認識によって判断すればよい。カテゴリ単語が抽出された場合、その少なくとも一部のカテゴリ単語について、カテゴリ単語に対応する単語をキーワードとしておけば、検索キーワードには処理を行なわずに通常のデータベースで用いるキーワード検索を用いることができるので、電子文書のデータと文書画像を同等に扱うことができるという効果がある。

【図面の簡単な説明】

【図1】　本発明の文書処理装置の第1の実施の形態を示す構成図である。

【図2】　本発明の文書処理装置の第1の実施の形態における類似文字分類部の処理の一例を示すフローチャートである。

【図3】　ペリフェラル特徴の説明図である。

【図4】　階層的クラスタリングの処理の一例を示すフローチャートである。

【図5】　クラスタリングの最適化処理の一例を示すフローチャートである。

【図6】　本発明の文書処理装置の第1の実施の形態における類似文字カテゴリテーブルの一例の説明図である。

【図7】　本発明の文書処理装置の第1の実施の形態における文字コード・カテゴリ対応テーブルの一例の説明図である。

【図8】　本発明の文書処理装置の第1の実施の形態における擬似文字認識部の処理の一例を示すフローチャートである。

【図9】　本発明の文書処理装置の第1の実施の形態における文字領域抽出結果の一例を示す説明図である。

【図10】　本発明の文書処理装置の第1の実施の形態における代表文字コード列への変換処理の一例を示すフローチャートである。

【図11】　本発明の文書処理装置の第1の実施の形態における代表文字コード列への変換処理の結果の一例を示す説明図である。

【図12】　本発明の文書処理装置の第1の実施の形態におけるbi-gramテーブルの一例の説明図である。

【図13】　本発明の文書処理装置の第1の実施の形態における代表文字コードテーブルの一例を示す説明図である。

【図14】　本発明の文書処理装置の第1の実施の形態における検索実行部の処理の一例を示すフローチャートである。

【図15】　本発明の文書処理装置の第1の実施の形態の第1の変形例における複数のカテゴリへの分類を許容した場合の文字コード・カテゴリ対応テーブルの一例の説明図である。

【図16】　本発明の文書処理装置の第1の実施の形態の第2の変形例における複数の文字切り出し解釈が存在する場合の切り出し位置の具体例を示す説明図である。

【図17】　本発明の文書処理装置の第1の実施の形態の第2の変形例における複数の文字切り出し解釈が存在

する場合の切り出された文字列の関係の説明図である。

【図18】　本発明の文書処理装置の第1の実施の形態の第2の変形例における複数の切り出し解釈を許容した場合の代表文字コードテーブルの一例の説明図である。

【図19】　本発明の文書処理装置の第1の実施の形態の第2の変形例における複数の切り出し解釈を許容した場合の代表文字コードテーブルの作成処理の一例を示すフローチャートである。

【図20】　本発明の文書処理装置の第1の実施の形態の第2の変形例における複数の切り出し解釈を許容した場合のｂｉ－ｇｒａｍテーブルの一例の説明図である。

【図21】　本発明の文書処理装置の第2の実施の形態を示す構成図である。

【図22】　本発明の文書処理装置の第2の実施の形態におけるカテゴリ単語辞書の一例の説明図である。

【図23】　本発明の文書処理装置の第2の実施の形態におけるカテゴリ単語辞書の別の例の説明図である。

【図24】　本発明の文書処理装置の第2の実施の形態におけるコード変換テーブルの一例の説明図である。

【図25】　本発明の文書処理装置の第2の実施の形態におけるカテゴリ単語検出部の動作の一例を示すフローチャートである。

【図26】　本発明の文書処理装置の第2の実施の形態におけるカテゴリ単語検出部の動作の一例を示すフローチャート（続き）である。

【図27】　本発明の文書処理装置の第2の実施の形態における品詞接続辞書の一例の説明図である。

【図28】　本発明の文書処理装置の第2の実施の形態における品詞接続関係の検証処理の一例を示すフローチャートである。

【図29】　代表文字コードの誤識別の一例の説明図である。

【図30】　本発明の文書処理装置の第2の実施の形態の第1の変形例におけるＮ＝２とした場合の代表文字コード列への変換の一例の説明図である。

【図31】　本発明の文書処理装置の第2の実施の形態における第2の変形例を示す構成図である。

【図32】　本発明の文書処理装置の第2の実施の形態の第2の変形例において詳細識別辞書を作成するために用いる特徴量の一例の説明図である。

【図33】　本発明の文書処理装置の第2の実施の形態の第2の変形例における詳細識別辞書の作成手順の一例を示すフローチャートである。

【図34】　本発明の文書処理装置の第2の実施の形態の第2の変形例における詳細識別辞書の一例の説明図である。

【図35】　本発明の文書処理装置の第2の実施の形態の第2の変形例における詳細識別部の処理の一例を示すフローチャートである。

【図36】　本発明の文書処理装置の第2の実施の形態の第3の変形例における切り出された文字列の関係の一例の説明図である。

【図37】　本発明の文書処理装置の第2の実施の形態の第3の変形例における切り出された文字列の関係の別の例の説明図である。

【図38】　本発明の文書処理装置の第2の実施の形態の第3の変形例における切り出された文字列の統合処理の一例を示すフローチャートである。

【図39】　本発明の文書処理装置の第2の実施の形態の第3の変形例における切り出された文字列の統合処理の一例を示すフローチャート（続き）である。

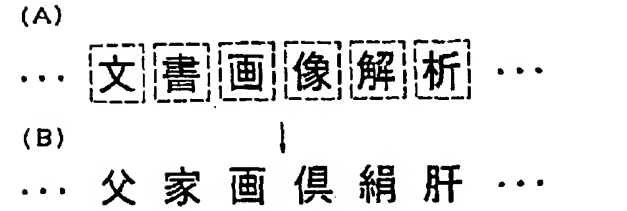
【符号の説明】

1…プロセッサ、2…表示装置、3…キーボード、4…マウス、5…スキャナ、6…プリンタ、7…外部記憶装置、11…類似文字分類部、12…擬似文字認識部、13…検索実行部、101…画像入力部、102…画像表示部、103…類似文字分類部、104…テキスト領域抽出部、105…擬似文字認識部、106…カテゴリ単語検出部、107…カテゴリ単語変換部、108…中央制御装置、109…記憶装置、110…詳細識別部、111…文字カテゴリ保持部、112…擬似文字認識結果記憶部、113…カテゴリ単語辞書、114…コード変換テーブル、115…詳細識別辞書。

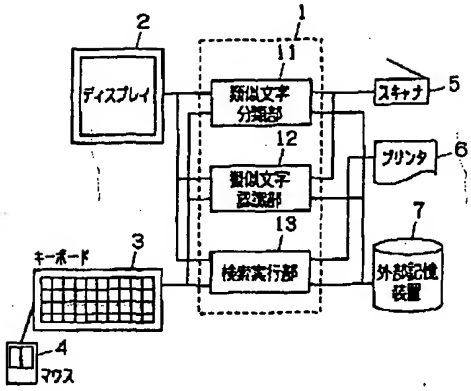
【図6】

代表文字	類似文字	代表ベクトル
亜	亜、並、疋、茎	(0.52, 0.91, …)
阿	阿、河	(0.01, 0.02, …)
囲	囲、困、固、圃、圏、国、団、…	(0.01, 0.01, …)
父	父、文、交	(0.23, 0.11, …)
家	家、賀、駕、要、茨、蚤、暮、…	(0.51, 0.02, …)
画	画、画、面	(0.01, 0.48, …)
倶	倶、仮、偽、協、像、徳、備、…	(0.25, 0.10, …)
絹	絹、旅、豚、賠、筒	(0.11, 0.09, …)
肝	肝、汗、析、祈、軒、新	(0.01, 0.04, …)
	⋮	

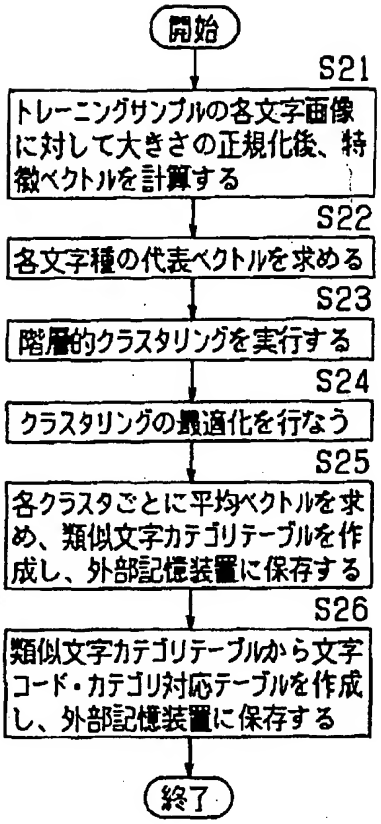
【図11】



【図1】



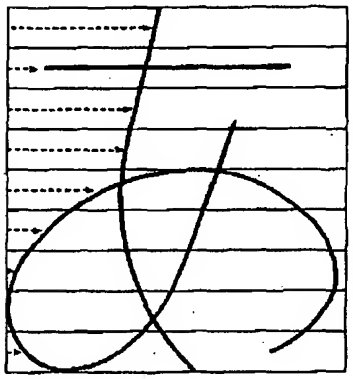
【図2】



【図7】

文字	カテゴリ代表文字
亜	亜
並	亜
直	亜
禁	亜
父	父
丈	父
文	父
家	家
貫	家
房	家
妻	家
書	家
⋮	

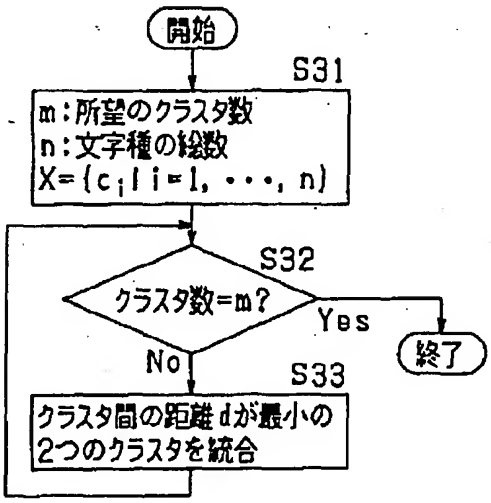
【図3】



【図13】

代表文字コード	画像上の位置
あ	(15,20,41,39)
⋮	
父	(120,340,43,42)
家	(165,341,41,43)
棚	(209,339,44,43)
倶	(253,340,41,42)
綱	(299,339,43,43)
肝	(343,339,42,42)
⋮	

【図4】



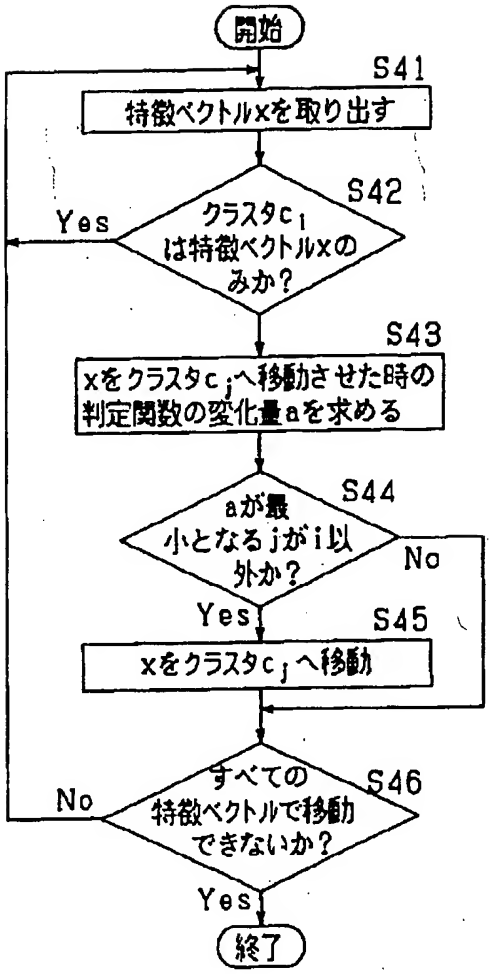
【図15】

文字	カテゴリ代表文字
亜	亜
並	亜、平
直	亜
禁	亜
父	父、交
丈	父
文	父、交
⋮	
書	家
⋮	
面	面
⋮	
倶	倶、場
⋮	

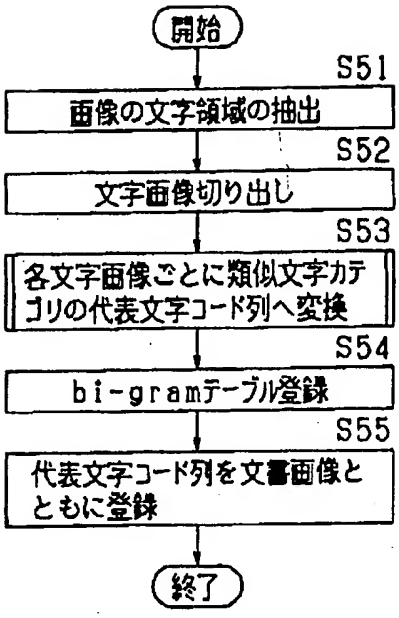
【図22】

代表文字コード列	品詞	文字単語
⋮	⋮	⋮
均衡	名詞	将来
⋮	⋮	⋮
芸道	名詞 名詞・サ変動詞語幹	実現
	名詞	差損
範用	名詞 名詞・サ変動詞語幹	構成
⋮	⋮	⋮
さ	下一段動詞語幹	き
	サ変動詞語尾・未然形	さ
⋮	⋮	⋮
れる	助動詞 下一段動詞語尾	れる
⋮	⋮	⋮

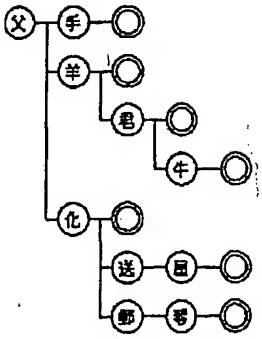
【図5】



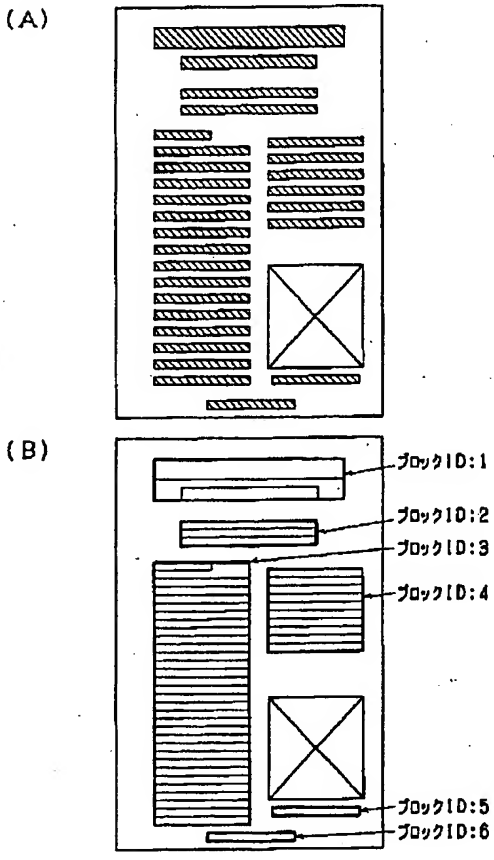
【図8】



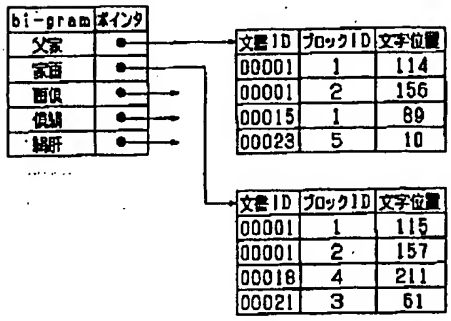
【図23】



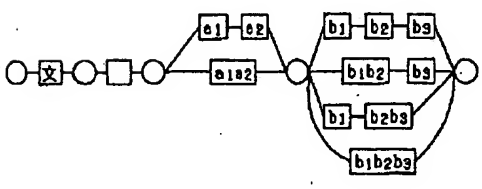
【図9】



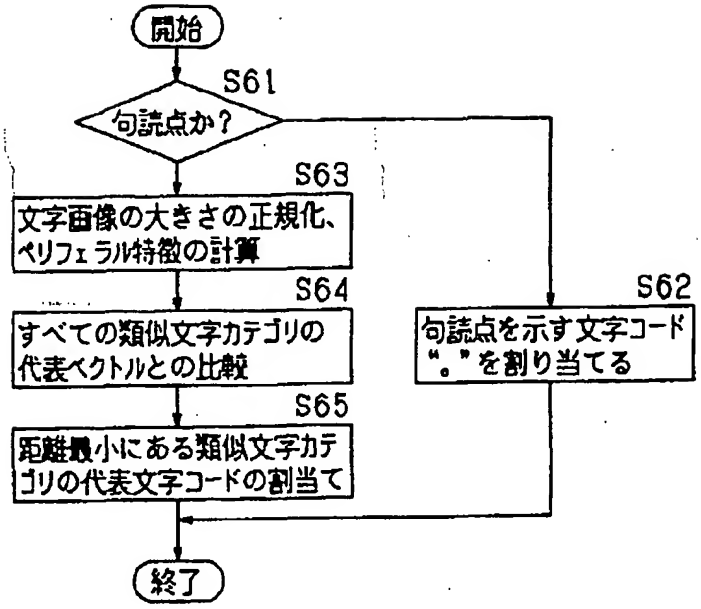
【図12】



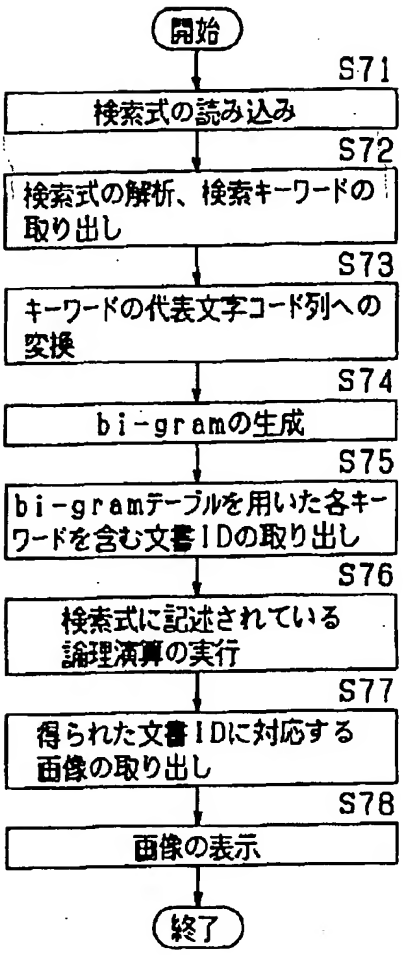
【図17】



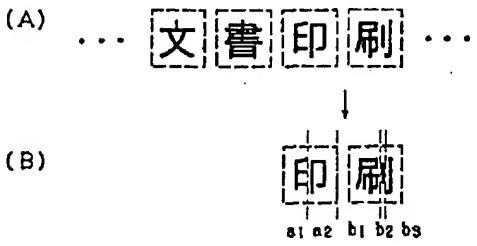
【図10】



【図14】



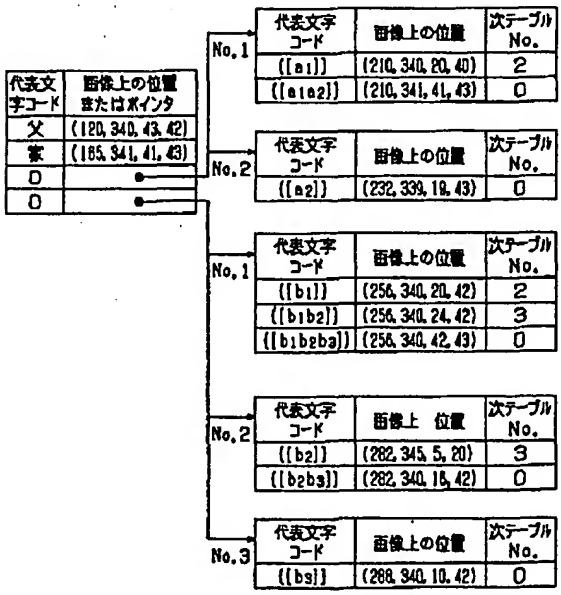
【図16】



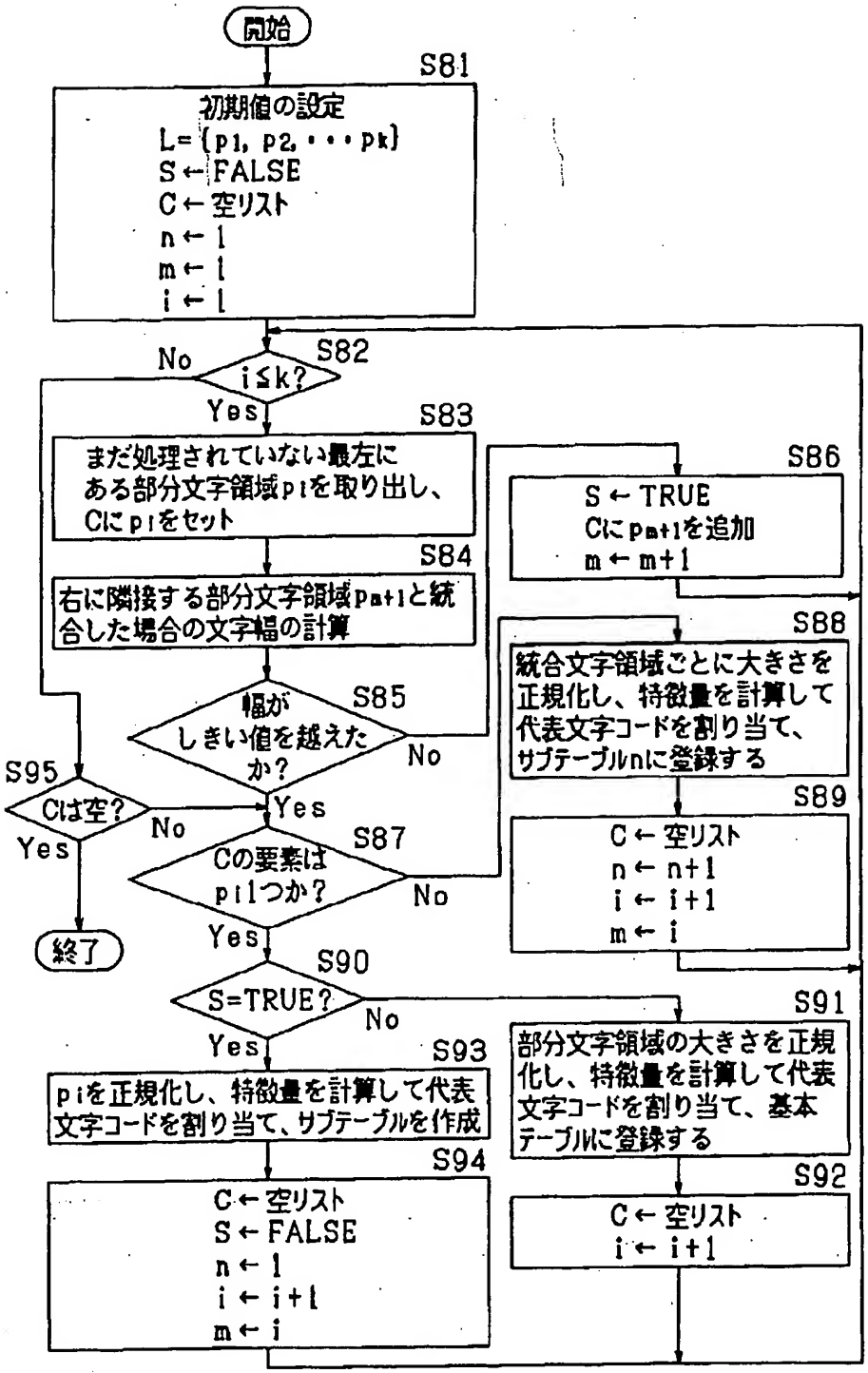
【図24】

代表文字コード列	文字単語	品詞
：	：	：
芸術	実現 差損 実現	名詞 名詞 名詞・サ変動詞語幹
：	：	：
稲用	構成 構成	名詞 名詞・サ変動詞語幹
：	：	：

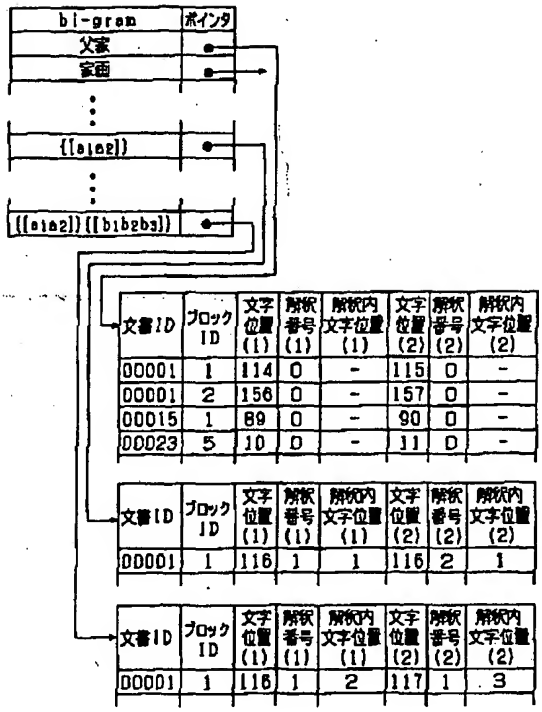
【図18】



【図19】


開始
S81
初期値の設定
L={p1, p2, ・・・pk}
S←FALSE
C←空リスト
n←1
m←1
i←1
No
S82
i≦k?
Yes
S83
まだ処理されていない最左にある部分文字領域piを取り出し、Cにpiをセット
S84
右に隣接する部分文字領域pm+1と統合した場合の文字幅の計算
幅がしきい値を越えたか?
S85
No
Yes
S86
S←TRUE
Cにpm+1を追加
m←m+1
S88
統合文字領域ごとに大きさを正規化し、特徴量を計算して代表文字コードを割り当て、サブテーブルnに登録する
S89
C←空リスト
n←n+1
i←i+1
m←i
S87
Cの要素はpi1つか?
No
Yes
S90
S=TRUE?
No
Yes
S91
部分文字領域の大きさを正規化し、特徴量を計算して代表文字コードを割り当て、基本テーブルに登録する
S92
C←空リスト
i←i+1
S93
piを正規化し、特徴量を計算して代表文字コードを割り当て、サブテーブルを作成
S94
C←空リスト
S←FALSE
n←1
i←i+1
m←i
S95
Cは空?
No
Yes
終了

【図20】

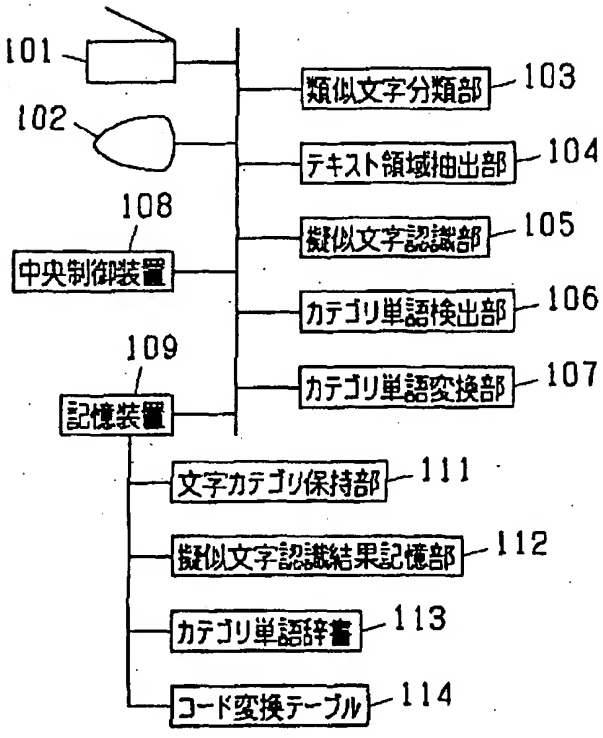


文書ID	ブロックID	文字位置(1)	解釈番号(1)	解釈内文字位置(1)	文字位置(2)	解釈番号(2)	解釈内文字位置(2)
00001	1	114	0	-	115	0	-
00001	2	156	0	-	157	0	-
00015	1	89	0	-	90	0	-
00023	5	10	0	-	11	0	-

文書ID	ブロックID	文字位置(1)	解釈番号(1)	解釈内文字位置(1)	文字位置(2)	解釈番号(2)	解釈内文字位置(2)
00001	1	116	1	1	116	2	1

文書ID	ブロックID	文字位置(1)	解釈番号(1)	解釈内文字位置(1)	文字位置(2)	解釈番号(2)	解釈内文字位置(2)
00001	1	116	1	2	117	1	3

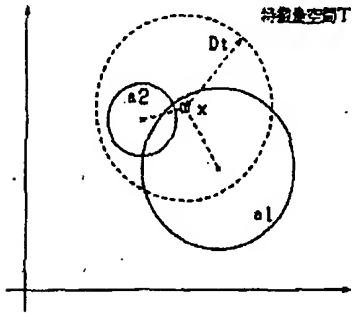
【図21】



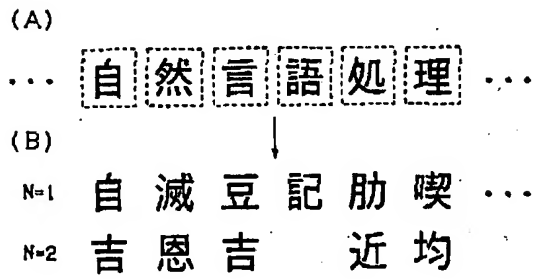
【図27】

	動詞活用語尾	サ変動詞語尾	助動詞「れる」	文頭特徴	…
名詞・サ変動詞語幹	0	1	0	0	…
名詞・形容動詞語幹	0	0	0	0	…
名詞	0	0	0	1	…
上一段活用動詞語幹	1	0	0	0	…
下一段活用動詞語幹	1	0	0	0	…
語尾・未然形	0	0	1	1	…
助動詞	0	0	0	1	…
⋮	⋮	⋮	⋮	⋮	⋮

【図29】



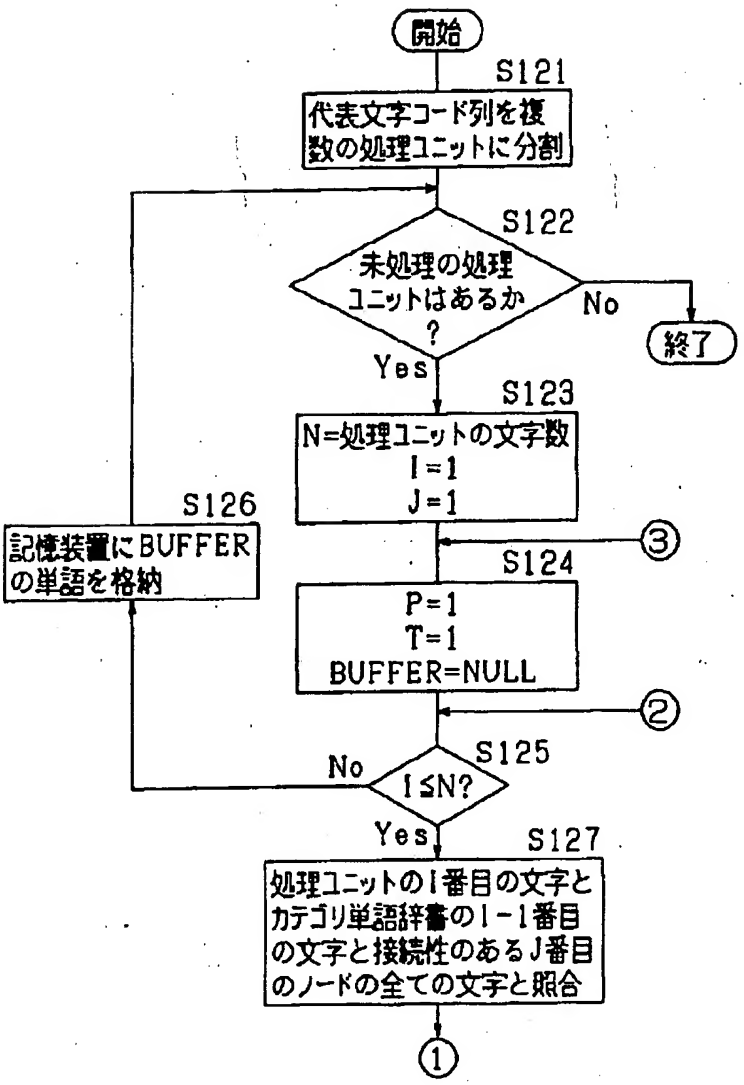
【図30】



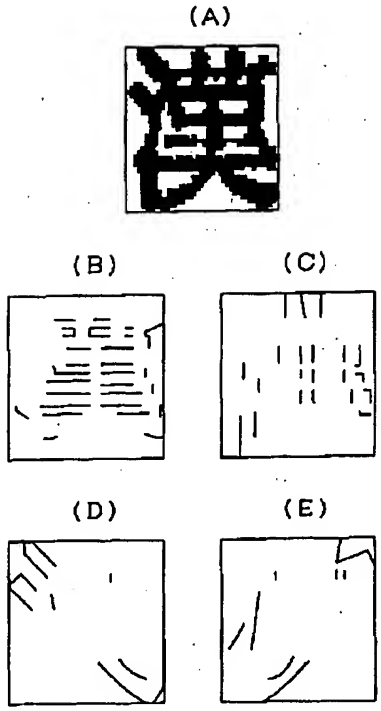
【図34】

代表文字	類似文字	詳細特徴ベクトル
亜	亜	(0.32, 0.54, 0.78, …)
	正	(0.36, 0.72, 0.29, …)
	正	(0.19, 0.21, 0.54, …)
	亞	(0.08, 0.58, 0.11, …)
	並	(0.28, 0.32, 0.27, …)
阿	阿	(0.65, 0.76, 0.92, …)
	何	(0.12, 0.32, 0.04, …)
	河	(0.43, 0.48, 0.57, …)
⋮	⋮	⋮
雨	雨	(0.25, 0.97, 0.86, …)
	而	(0.34, 0.33, 0.54, …)
	両	(0.12, 0.54, 0.34, …)
	丙	(0.77, 0.64, 0.68, …)
⋮	⋮	⋮

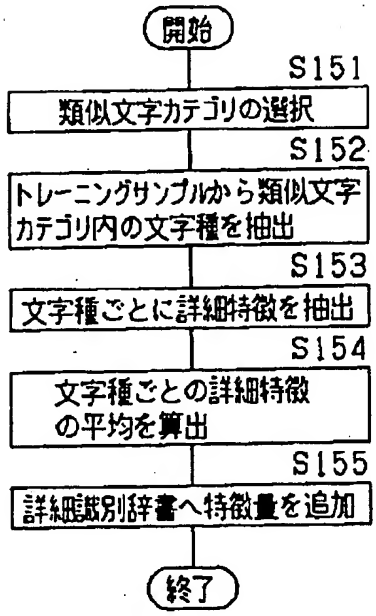
【図25】



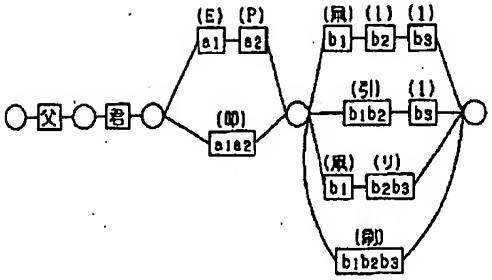
【図32】



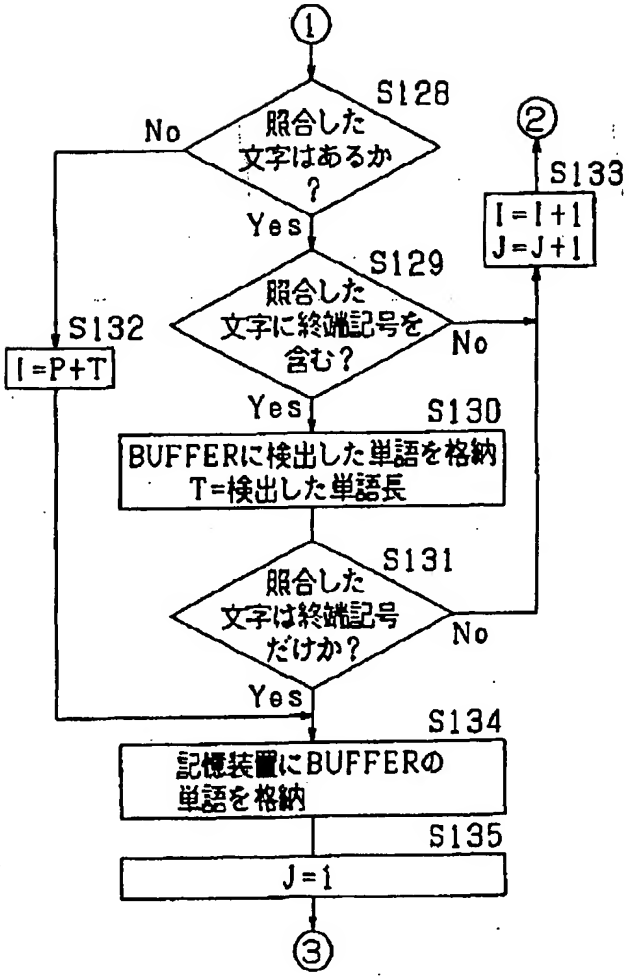
【図33】



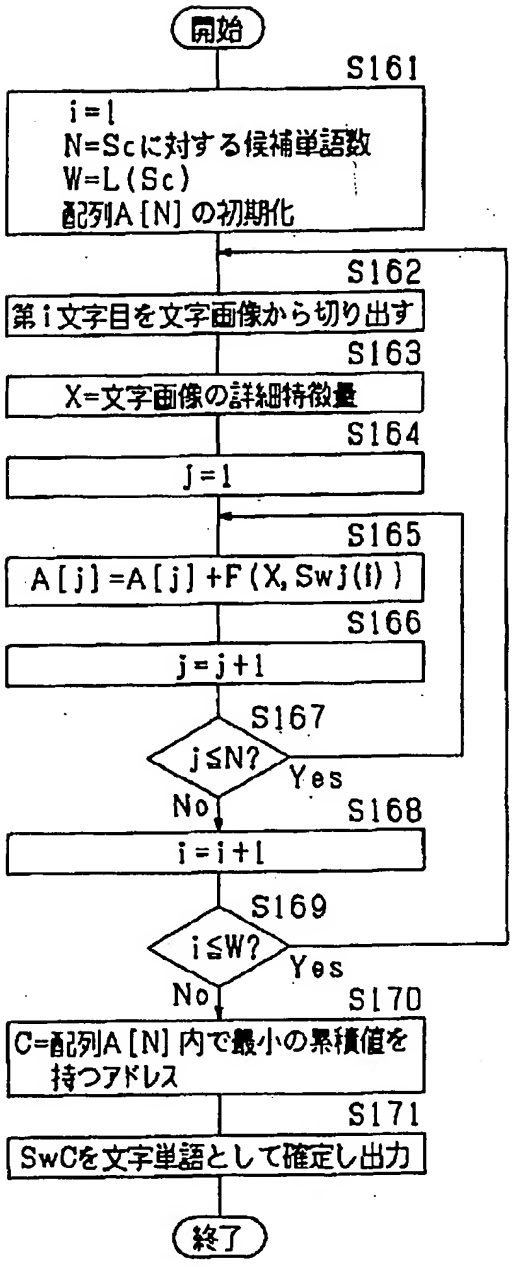
【図36】



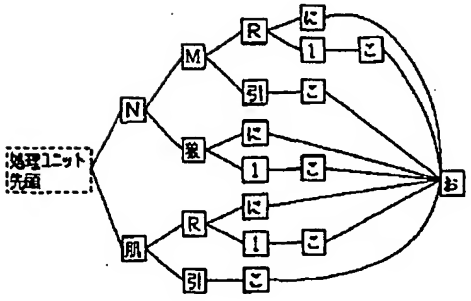
【図26】


①
S128
照合した文字はあるか？
No
Yes
S129
照合した文字に終端記号を含む？
No
Yes
S130
BUFFERに検出した単語を格納
T=検出した単語長
S131
照合した文字は終端記号だけか？
No
Yes
S132
I=P+T
②
S133
I=I+1
J=J+1
S134
記憶装置にBUFFERの単語を格納
S135
J=1
③

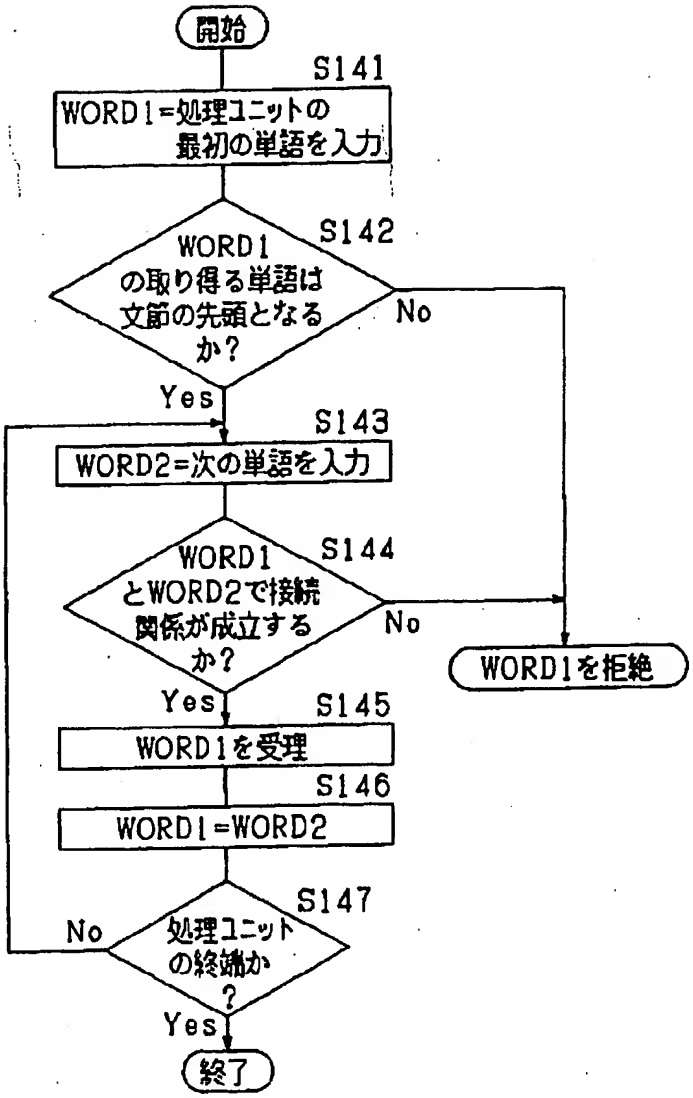
【図35】


開始
S161
i=1
N=Scに対する候補単語数
W=L(Sc)
配列A[N]の初期化
S162
第i文字目を文字画像から切り出す
S163
X=文字画像の詳細特徴量
S164
J=1
S165
A[j]=A[j]+F(X,Swj(i))
S166
j=j+1
S167
j≦N?
Yes
No
S168
i=i+1
S169
i≦W?
Yes
No
S170
C=配列A[N]内で最小の累積値を持つアドレス
S171
SwCを文字単語として確定し出力
終了

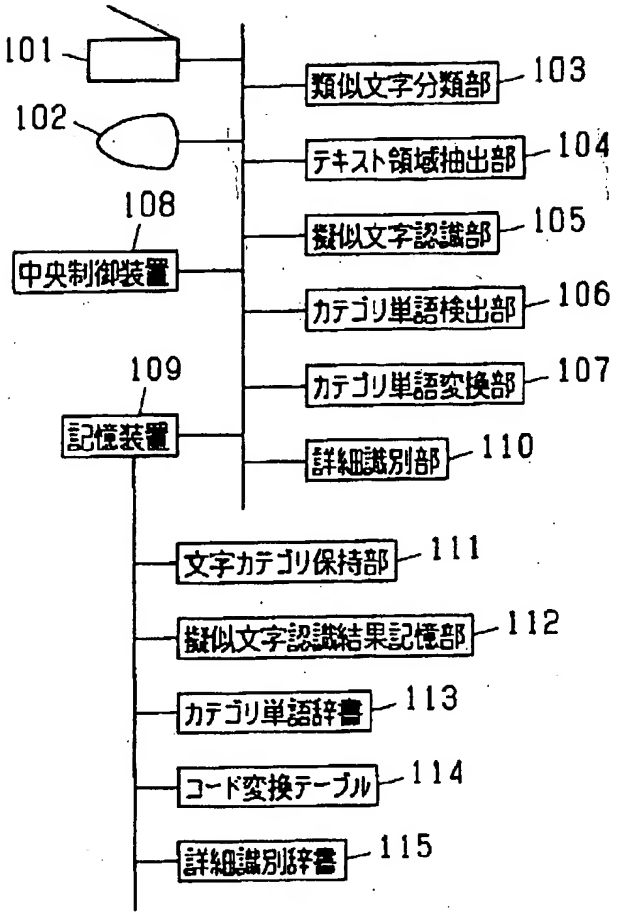
【図37】


処理ユニット
先頭
N
M
R
に
1
こ
引
こ
に
1
こ
R
に
1
こ
引
こ
お

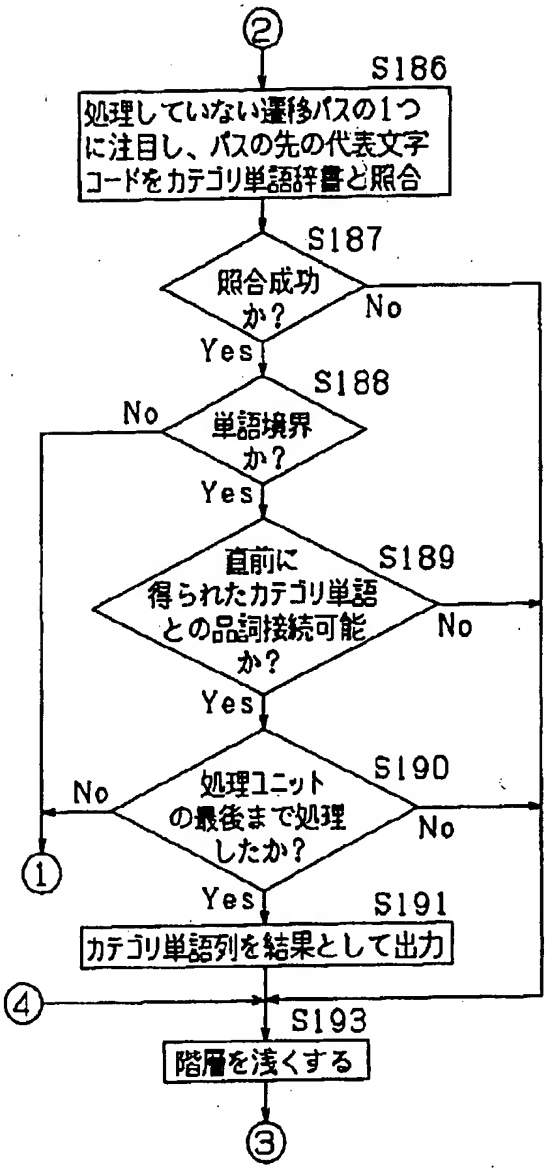
【図28】


開始
S141
WORD1=処理ユニットの最初の単語を入力
S142
WORD1の取り得る単語は文節の先頭となるか?
No
Yes
S143
WORD2=次の単語を入力
S144
WORD1とWORD2で接続関係が成立するか?
No
Yes
WORD1を拒絶
S145
WORD1を受理
S146
WORD1=WORD2
S147
No
処理ユニットの終端か?
Yes
終了

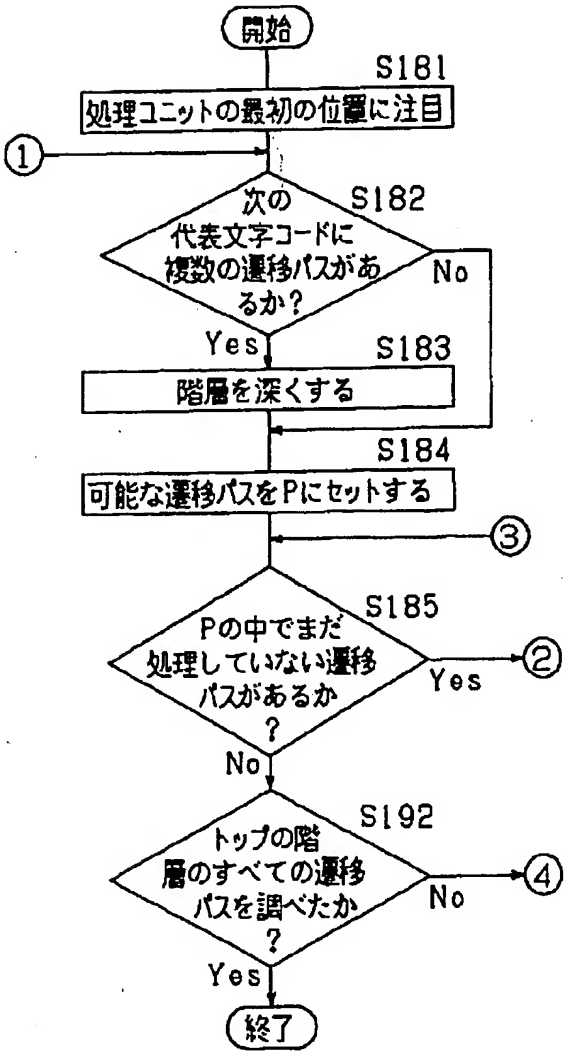
【図31】


101
102
103
類似文字分類部
104
テキスト領域抽出部
108
中央制御装置
105
擬似文字認識部
106
カテゴリ単語検出部
109
107
カテゴリ単語変換部
記憶装置
110
詳細識別部
111
文字カテゴリ保持部
112
擬似文字認識結果記憶部
113
カテゴリ単語辞書
114
コード変換テーブル
115
詳細識別辞書

【図39】


②
S186
処理していない遷移パスの1つに注目し、パスの先の代表文字コードをカテゴリ単語辞書と照合
S187
照合成功か？
No
Yes
S188
No
単語境界か？
Yes
S189
直前に得られたカテゴリ単語との品詞接続可能か？
No
Yes
S190
No
処理ユニットの最後まで処理したか？
No
Yes
①
S191
カテゴリ単語列を結果として出力
④
S193
階層を浅くする
③

【図38】


開始
S181
処理ユニットの最初の位置に注目
①
次の代表文字コードに複数の遷移パスがあるか？
S182
No
Yes
S183
階層を深くする
S184
可能な遷移パスをPにセットする
③
Pの中でまだ処理していない遷移パスがあるか？
S185
Yes
②
No
トップの階層のすべての遷移パスを調べたか？
S192
No
④
Yes
終了